Chapter 3

On the Banks of Shodhganga: Analysis of the Academic Genealogy Graph of an Indian ETD Repository

Academic genealogy construction of a discipline is fraught with difficulties, particularly those caused by data sparsity and researcher name ambiguity. Some common data fields whose sparsity have a marked effect on the completeness of the genealogy network are the exhaustive list of protégés of a researcher, names of the advisor of a researcher, year of graduation, details of institutional affiliation of the researcher, and the discipline. Researcher name ambiguity occurs because it is not easy to resolve whether two identical researcher names refer to the same individual or not, and whether two different names refer to the same individual. Given the scare information available in the genealogy records, it is challenging to disambiguate researcher names with high accuracy. Although there are prominent initiatives like the Mathematical Genealogy Project[1] and Academic Tree[2] to build high-quality academic genealogy networks (AGNs), they do not cover all researchers even in their respective disciplines, and they do not represent researchers from all

[1]https://www.genealogy.math.ndsu.nodak.edu/
[2]https://academictree.org/

countries equally well [41].

Given the relatively low presence of researchers from the Indian universities in the above global databases, we are motivated to explore the academic genealogy of only the researchers who have submitted their doctoral dissertations in India. Shodhganga ('Shodh' means research in Hindi, and 'Ganga' is a large river in India) is a repository of doctoral theses and dissertations submitted to Indian Universities, and houses more than $300,000$ theses as on April 9, 2021. The metadata associated with each thesis includes the name of the doctoral researcher, name of their advisor, department and university where the thesis was submitted, date of submission, dissertation title, and keywords. In this paper, we build an academic genealogy network using the metadata available in Shodhganga. We undertake a detailed analysis of this AGN and unravel several patterns in the higher educational sphere of India over the years. Our contributions are as follows.

1. We construct an AGN from the Shodhganga repository after disambiguating the names of researchers that appear in the repository. The AGN contains more than 256K researchers and 20K advisor-advisee relationships.

2. We analyze the largest connected component of researchers in the Shodhganga-AGN. It contains 1356 researchers and 1437 advisor-advisee relationships. The component is dominated by the researchers from science and affiliated primarily with three institutions.

3. We perform a detailed analysis of the AGN using network-based metrics at the researcher, institute, and subject levels. We study the temporal evolution of different features over 10-year intervals, like the number of researchers graduated and the distance between institute ranks. At the subject level, we study the importance of different subjects (based on DDC subject classification) associated with theses over time.

4. We study the collaboration patterns of researchers in Shodhganga-AGN considering three distinct subgraph structures.

3.1 Construction of AGN from Shodhganga Repository

The Shodhganga repository consists of multidisciplinary theses from different universities and institutions in India. The total number of theses that we extracted from Shodhganga for this study is 205235, which contains theses from 1883 to 2019 (thesis submission date). Each thesis record consists of multiple attributes. The available attributes include advisor name, advisee name, publisher institution, publisher department, title, date submitted, DDC subject classification, abstract, and alternative title. However, not all records contain values for all attributes. In Fig. 3.1, we have shown the statistics of available information for attributes.

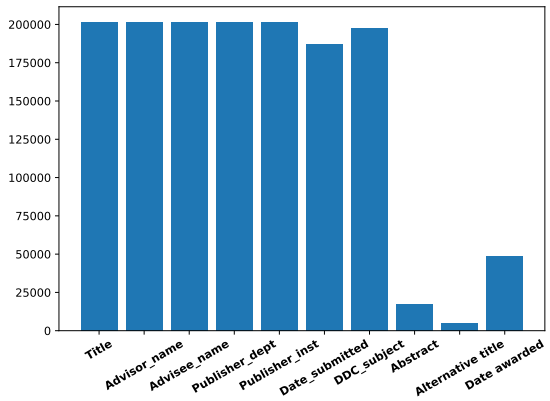


Figure 3.1: Available attribute values of theses in Shodhganga repository

3.1.1 Growth of Dissertation in India

There is no literature that provides data on the number of PhD graduates in India as of 2019. To understand the actual growth of dissertations in India over time, we retrieved data from the "All India Survey on Higher Education (AISHE)" website[3] (survey conducted by the "Ministry Of Education" in Higher Education) which is shown in Table 3.1. As enrolment data for PhD was only accessible for six years, we extracted the total number of students enrolled in higher education between 1950 and 2020 over 10 year interval which was available on

[3]https://aishe.gov.in/aishe/gotoAisheReports

Year	PhD enrolled	PhD Awarded	PhD enrolment percentage
2019-2020	202550	38986	$\leq 0.5\%$
2018-2019	169170	40813	$< 0.5\%$
2017-2018	161412	34400	$< 0.5\%$
2016-2017	141037	28779	$< 0.4\%$
2014-2015	117301	21830	$< 0.34\%$
2012-2013	95425	23630	$< 0.4\%$

Table 3.1: All India Survey on Higher Education ($< 0.5\%$ means less than 0.5 percentage students have enrolled in PhD out of total students enrolled in Higher Education)

the University Grants Commission (UGC) website[4] (Higher Education Growth Data). According to the data in Table 3.1, less than or equal to 0.5% of total enrolled students in higher education were pursuing a PhD. Fig. 3.2 displays the number of anticipated PhD students enrolled (0.5% of total students enrolled) from 1950 to 2020 over a 10-year interval. The approximated total number of PhD students between 1950 and 2020 is 540420 (aggregated PhD students in 70-years over 10 year intervals). Based on approximated PhD enrolments the number of available theses in Shodhganga until 2019 accounts for 38%. The dropout rate in PhD programs in India has not been taken into account. After taking the drop rate into account[5], we estimate that there are about 40% of theses in Shodhganga.

Based on also the PhD awarded information available in Table 3.1 between 2012-2020, which aggregates to 188438 and the number of PhD theses listed in the Shodhganga repository after 2012 (i.e., 72825, Shodhganga includes theses between 2012-2019 and only 33 theses available between 2019-2020), which accounts for 38%. The percentage of available theses in the repository between 2012 and 2019 will rise (i.e., account for 48%) when only the PhD awarded information available between 2012 and 2019 from Table 3.1 is taken into account.

3.1.2 Data Model for the AGN

AGN is a directed acyclic graph $G(V, E)$ in which V represents researchers and E represents advisor-advisee relationships. A directed acyclic graph is directed

[4] `https://www.ugc.ac.in/pdfnews/8604044Higher-Education-Brochure.pdf`
[5] Dropout rate is not known, but we assumed it to be more than 2%.

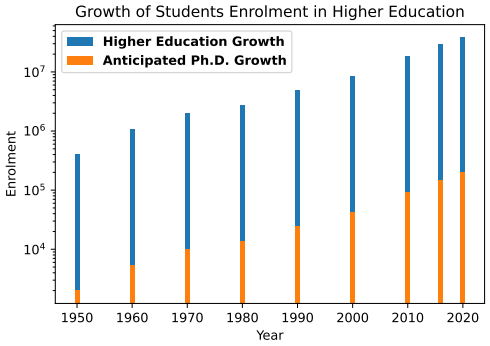


Figure 3.2: Growth of students enrolment in higher education and anticipated PhD enrolment in India

graph with no directed cycles.

The data model for the AGN is shown in Fig. 3.3. In this data model, "Person", "Publication", and "Institute" refer to concepts and "MENTOR", "ADVISED", "RESEARCHED", and "PUBLISHEDBY" refer to relationships. The data model has been implemented in an open-source graph database Neo4j[6]. The "Person" class or concept has been used to model the researchers. The concept "Publication" has been used to typify the thesis and "Institute" is the class of the degree awarding institutions.

The advisor and advisee ("Person" Class) information in theses are related with "MENTOR" relation, the advisor and advised theses are related with "ADVISED" relation, the advisee and his defended thesis are related with "RESEARCHED" relation, and theses ("Publication" Class) and degree awarding institute ("Institute" Class) are related with "PUBLISHEDBY" relation. All the available components of the data model are extracted from the thesis metadata available in the Shodhganga repository (i.e., Advisor-Advisee, Institute, Thesis Title etc.).

To construct AGN, we first extracted data from Shodhganga in tabular format, with each record/row representing thesis metadata. The thesis' metadata comprises information about the thesis's advisor and advisee. If a thesis has several advisors, we replicated the records as many times as there were advisors, with the exception of the advisor's name, which varies across rows. To construct the AGN, we must disambiguate the identities of the researchers. Since the majority of names

[6]https://neo4j.com/

in Shodhganga are not included in well-known external knowledge bases such as Wikipedia, we can only do the disambiguation using information within the dataset. Firstly, we assigned each researcher (advisor or advisee) with different index in the dataset and then after applying successive algorithms explained below we merge the indices (i.e. researchers assigned with different index but potentially referring to same researcher).

In the subsequent subsections, we discuss our proposed strategy for disambiguating the names of researchers.

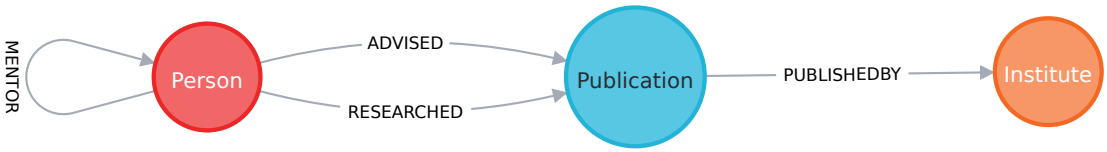


Figure 3.3: Data Model for Shodhganga-AGN

3.1.3 Dataset Disambiguation Approach

In Fig. 3.4 we have shown the step-by-step procedure we followed to disambiguate the researchers in Shodhganga dataset. We have made an assumption that the thesis publisher *department* and *institute* are the same for the advisor and the advisee related to the thesis. Initially, we considered researcher names

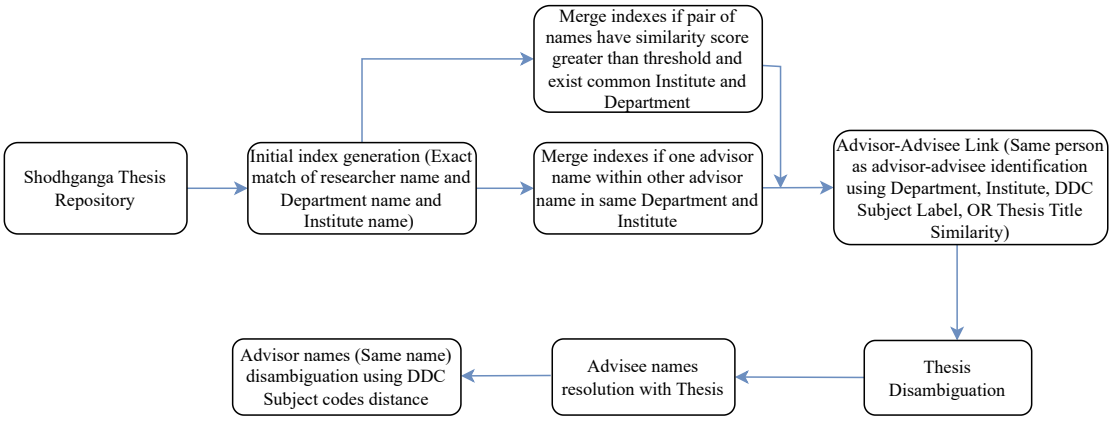


Figure 3.4: Researchers disambiguation approach to create an academic genealogy network

that are exactly same (exact string match) and belong to the same department and institute, as referring to the same person (*Initial Index Merging*). After that, we developed Algorithm 1 (Fig. 3.5) which uses sound-based, token-based, and

character-based information from a pair of names to provide a similarity score. If a pair of names has a similarity score greater than a threshold value and there is at least one common institute and department between them, then those names are considered the same and re-indexed with same index (This handles the case of different structural variations in names referring to the same individual/researcher in the dataset). Some of the variations of the names which are disambiguated with Algorithm 1 (Flowchart in Fig. 3.5) are shown below (Names in each row potentially refer to a single individual):

```
"Gudadhe, S. V."; "Gudadhe S. V."; "Gudhade S. V."
"Ajith Kumar, N."; "Kumar, N. Ajith"; "Ajithkumar, N."
"Singh, Meenakhsi"; "Singh, Menakshi"; "Sing, Meenakshi"
"Rathod, Urvashi"; "Urvashi, Rathod"
"Kumar, Sanjeev"; "Kumar, Sanjiv"
```

In the dataset, there are names which refer to the same person but have different variations that are difficult to capture with a distance-based metric (similarity metric); in these cases, if a name is embedded within another name (e.g., an advisor name was found to be sub-part of the other advisor name) and there is only one candidate key (name) which covers that name within the department, then the two instances are assumed to refer to the same individual, hence assigned the same index. For example, each row below contains 2 advisor names which our algorithm infers to be variations of the same name:

```
"Joshi, Manoj"; "Joshi, Manojbhai"
"Jani, Balvant"; "Jani, Balvantbhai S."
"Joshi, Daxa"; "Joshi, Daxaben"
"Vimala"; "Yadav, Vimala"
```

To combine names that appear as advisee and as advisor in the dataset, we have developed a rule-based Algorithm 2 (Flowchart in Appendix Fig. 7.1), which leverages the similarity between their respective department names, institute names, DDC subject information of the associated dissertations, and thesis titles to infer if they refer to the same individual. We have used the `spaCy`[7] pre-trained model to obtain the similarity score across departments, institutes, DDC subjects, and theses

[7] `https://spacy.io`

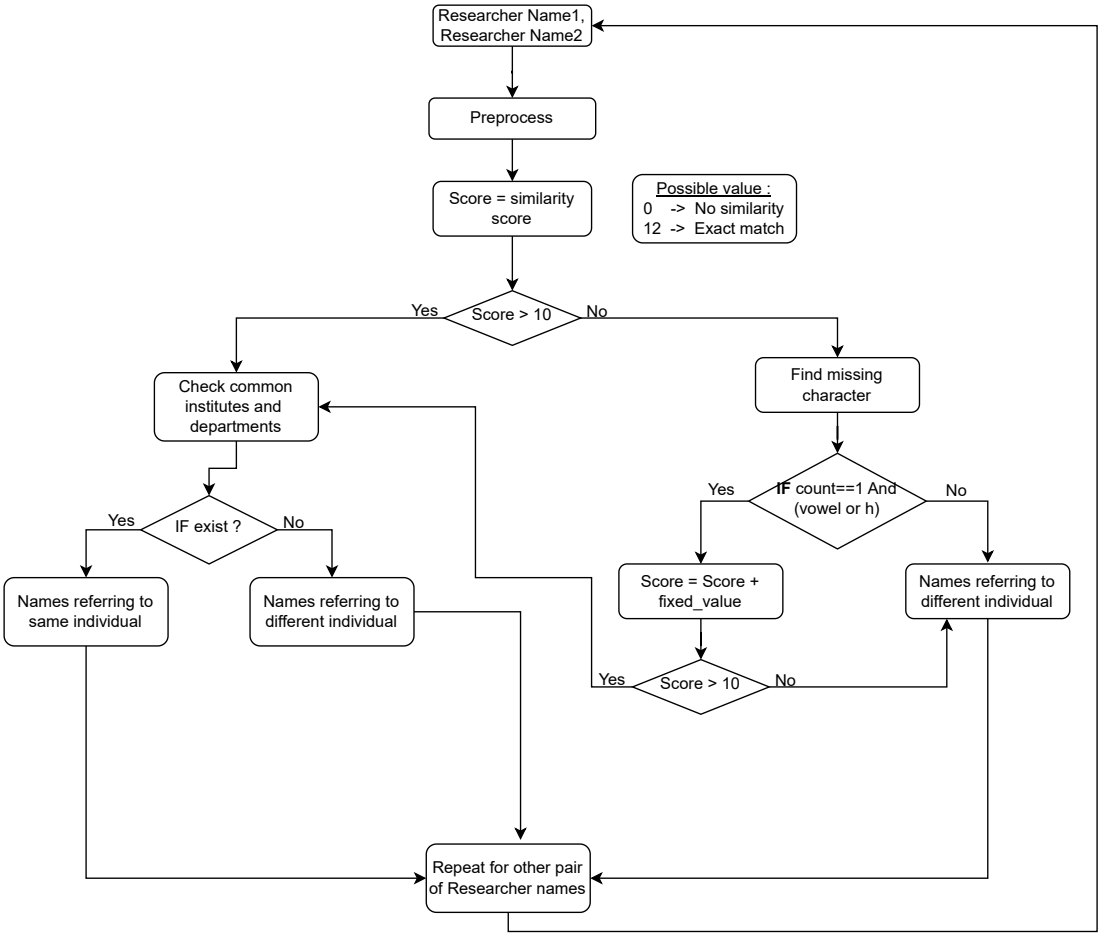


Figure 3.5: Algorithm 1 - Identify researchers which are referring to same individual. In some case a sequence of character is mapped to single character or a null character. For those case we added fixed value to the similarity score.

title (to capture the semantic/syntactic similarity; similar names across different institutes or departments are combined if they are referring to the same individual). The similarity value is smaller if many attributes are matched across the records of the individual as advisor and advisee both, and as the number of matched attributes across records decreases, the similarity score increases (the threshold value[8] decreases as the number of attributes matched between the researchers records increases). For example, the researcher's names (as advisor and advisee), which potentially refer to the same individual are shown below:

```
As student: Jose, Joseph -- department of botany, University of Calcutta.
As advisor: Jose, Joseph -- department of sacred heart college.(thevara)
            department of botany, Mahatma Gandhi University.
As student: Bandhu, Tarlok -- department of education,
Himachal Pradesh University.
As advisor: Bandhu, Tarlok -- department of education, Panjab University.
```

To disambiguate the theses title, we first relied on exact string matching. Then to further disambiguate theses title which are similar but having syntactic variations (missing character, words, etc.), we have developed an algorithm similar to Algorithm 1 (Fig. 3.5). The Shodhganga repository contains theses in multiple regional languages. Therefore it was difficult to disambiguate them properly. Our focus while disambiguating the theses title was to only merge a pair of titles if the score returned by the similarity algorithm is very high (> 0.99). Note that our aim in disambiguating theses title was to improve the disambiguation of researchers name in the repository. *Advisee name disambiguation:* There are multiple records where the same advisee name appears but with different theses title. We assumed that a single thesis title corresponds to a single student. Based on that assumption, first we have indexed the theses title and then used the thesis index with thesis submission date for advisee name disambiguation. There were multiple common advisee names in the Shodhganga repository which have been assigned to same index initially. Advisees who have similar names and same index but have different thesis id, after applying Algorithm 3 and Algorithm 4 (Appendix Fig. 7.2, 7.3), are assigned a different index (i.e. they are now resolved to refer to different individuals). We have also combined advisee names which have the same thesis index, i.e. having same thesis title (referring to same advisee) but due to structural variation[9] in the advisee names they have not been indexed with same index

[8]In Algorithm 2 the the threshold value is dynamic

[9]Structural variation refers to characters missing, longer-shorter version of names, etc.

previously (duplicate records are available with variation in researcher names in the digital repository), are re-assigned with same index. The algorithms used for advisees disambiguation are explained with flowchart in Appendix Fig. 7.2, 7.3. A few names disambiguated by the algorithms are:

```
Student Name(Disambiguated Index); Department; University(Advisor Name); Title.

Kumar, Rajesh(279598); department of chemistry; Himachal Pradesh University
(Kalia, S. B.); "Physicochemical studies on some novel carbamate complexes of
copper_ii_ and zinc_ii_ with potential insecticidal activity".

Kumar, Rajesh(285560); department of chemistry; Himachal Pradesh University
(Blokhra,R.L.); "Ultrasonic studies of some solutions involving isopropyl
alcohol".

Kumar, Rajesh(121132); school of social sciences;Jawaharlal Nehru University
(Gandhi,J.S.); "Fluvial processes in lower Rapti river basin: a case study of
impacts on arable land".

Kumar, Rajesh(282352); school of social sciences; Jawaharlal Nehru University
(Sharma, Milap Chand); "Drug addiction and its relapse: a sociological study of
 drug addicts drawn from some de-addiction centres of Delhi".

Initially advisees having same names across institute and department are
assigned with same index even if they are referring to different individual,
but after disambiguation assigned with different indices.
```

```
Student Name; Initial Student Index; Final Student Index; Department;
University; Title; Initial Thesis Index; Final Thesis Index.

Karavadra Nehal, S.; 279598; 279598; department of education; Saurashtra
University; "Construction and standardization of scientific attitude and
scientific aptitude test"; T53420; T53420.

Karavadra, Nehal Sukabhai; 180849; 279598; department of education; Saurashtra
 University; "Construction and standardization scientific attitude and
  scientific aptitude test"; T99003; T53420.
```

```
Roopesh Kumar, T.; 263641; 263641; department of geo-engineering;
Andhra University; "Hieraarchial hybrid SVM method for joint spatial and
spectral classification of remotely sensed data"; T148565; T148565.

Kumar, Tamma Roopesh; 263642; 263641; department of geo-engineering;
Andhra University; "Hierarchical hybrid SVM method for joint spatial and
spectral classification of remotely sensed data"; T117333; T148565.

Based on the disambiguated thesis title index, the advisees' name which are
initially referring to different people are combined to point to same
individual.
```

The department names are sometimes very noisy and include the institute name or combination of the institute name and the department name, along with various other variations. Therefore, it was difficult to disambiguate the department names. That is the reason that many of the researchers (advisors) who refer to the same individual have been assigned different indices by previous approaches. In order to deal with this problem, we have developed Algorithm 5 (Flowchart in Fig. 3.6) that uses the absolute distance between the DDC subject codes. If the distance between the DDC codes are within a predefined threshold limit and the researchers' names are similar and they belong to the same institute, then they are re-indexed with the same index. For example,

```
Advisor Name;          Department Name;                University Name.
Adhya, T. K.(8438);    department of botany;           Utkal University.
Adhya, T. K.(12274);   department of microbiology;     Utkal University.
Adhya, T. K.(86019);   department of zoology;          Utkal University.

All the indices (8438, 12274, 86019) referring to different advisors
(previously) are merged to 86019 after disambigutaion.
```

In short, we considered various algorithms with available columns in the dataset (Title, Submission date, Institute, Department, DDC subject code, etc.) for the disambiguation of researcher names in the Shodhganga dataset. We also manually corrected some of the names. There were many duplicate records in the dataset

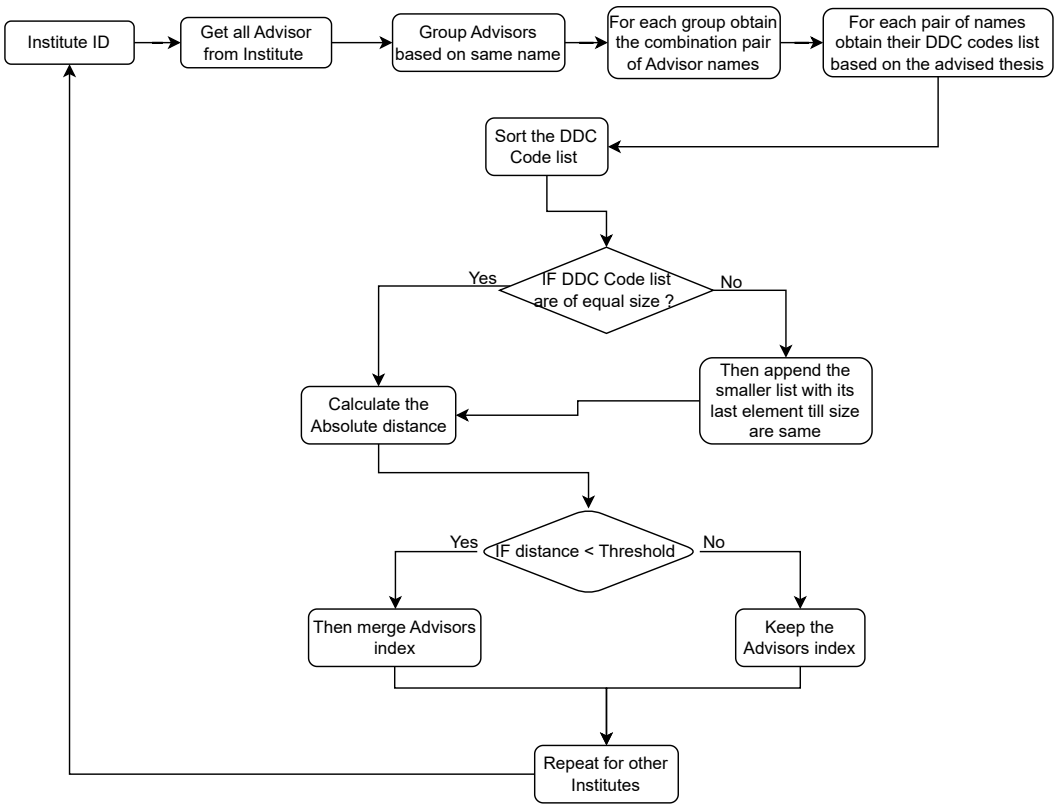


Figure 3.6: Algorithm 5 - Merge advisors having same name within a institute, if absolute distance between DDC codes list is less than threshold value

that we removed.

Some of the limitations of the disambiguation approach are that the researchers are not well disambiguated across institutes due to a lack of information in the dataset. Also, if the length difference between two names is very large and one of the names is not included within the other, they may not be disambiguated. If a person has completely different names, such as “Mahi Singh” and an alias name “Mahendra Singh” (synonym), the names will not be disambiguated to refer to the same person. There is a chance that some of the names will be incorrectly disambiguated because the other fields, such as “Thesis” and “Department”, are incorrectly disambiguated. However, whenever we used an inference-based approach to disambiguate names based on other fields, we ensured that the number of correctly disambiguated names is greater than the number of incorrectly disambiguated names (manually verified with random selection). We attempted to keep our disambiguation algorithm’s precision as high as possible. The reason for keeping precision high is that it will

help us in only merging the index of the names if they are referring to the same person; otherwise names will be considered different and have different indexes. Initially, we assumed the advisor/advisee names in all published theses (records) to be different.

The above approaches are generic and applicable to any English dissertation database based on the availability of similar thesis metadata. It could also be further extended based on availability of other external source of information for researchers/theses.

3.2 Characterization of the AGN

Following the disambiguation of researchers in the Shodhganga thesis repository, we created an academic genealogy graph called Shodhganga-AGN[10]. There are 256725 researchers and 201250 advisor-advisee relationships in the academic genealogy graph. We used Gephi[11] force-directed layout (ForceAtlas2) with parameters stronger gravity and approximate repulsion, and OpenOrd layout (with default parameter) to show the influential researchers of the Shodhganga-AGN subgraph in Fig. 3.8. The size of the nodes in the graph is proportional to the fecundity value (node size scaled based on degree value), and they are colored based on the DDC subjects. We have annotated a researcher with the most generic DDC subject code value (i.e., 300, 310, 390 $\implies$ 300) if he/she belongs to multiple DDC subject codes (based on his/her published or advised thesis).

Given a directed graph G, a weakly connected component (WCC) is a subgraph of the original graph where all vertices are connected to each other by some path, ignoring the direction of edges. The size distribution of weakly connected components (WCC) of the AGN is depicted in Fig. 3.7. The total number of components is 56,935, with 7% of components having a size greater than or equal to 10, and 52% having a size of 2. As shown in Fig. 3.8, the components with a size greater than or equal to 10 account for 46% of the edges (93559 out of 201250)

[10]https://github.com/Djasingh/Shodhganga-AGN
[11]https://gephi.org/

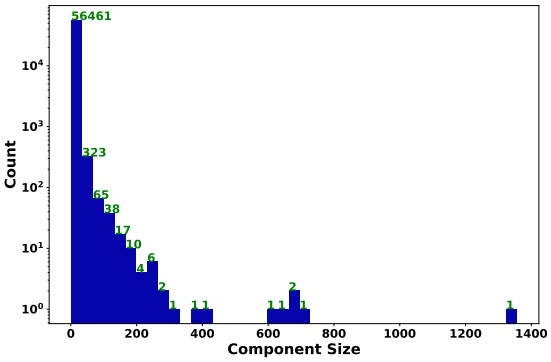


Figure 3.7: Size distribution of weakly connected components in the AGN.

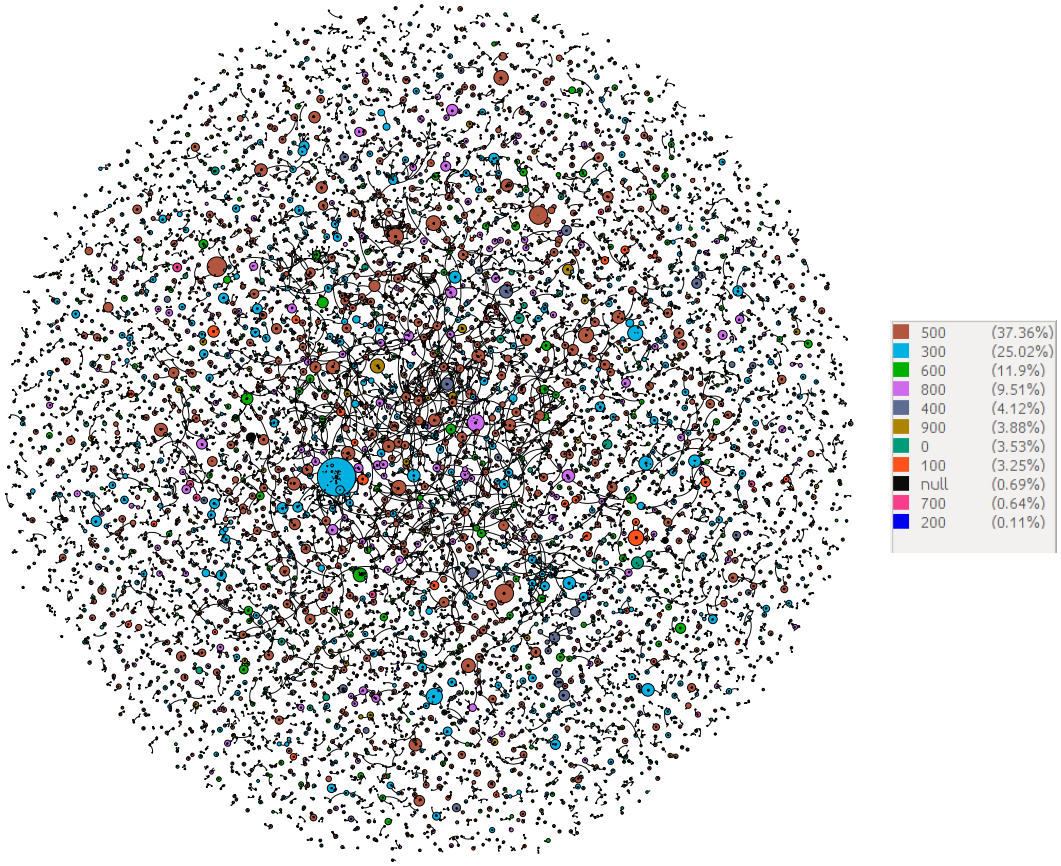


Figure 3.8: The overview of influential researchers across DDC subjects in Shodhganga-AGN subgraph. The subgraph consists of all the weakly connected components of size $\geq$ 10 which account for 46% of the edges and 37% of the nodes in the Shodhganga-AGN. In the graph node size is proportional to fecundity value, color represents DDC subjects, percentage value represents the proportion of researchers belonging to the respective DDC subjects. The graph is created with Gephi force directed layout (ForceAtlas2) with property stronger gravity and approximate repulsion, zoom in to visualize the image clearly (check Appendix Table 7.10 for DDC code names).

and 37% of the nodes (96514 out of 256725) in the AGN. In Graph Fig. 3.8, the largest components attracted towards center while smaller components lies along

the borders.

3.3 Study of the Largest Component

We have shown the largest component of the Shodhganga-AGN in Fig. 3.9, which consists of 1356 researchers and 1437 advisor-advisee relationships. The researchers name size (nodes label) are proportional to the researchers' fecundity values (out-degree). We used Gephi force-directed (ForceAtlas2) layout with the property prevent overlap and expansion layout to generate the graph visualization shown in Fig. 3.9. The distribution of researchers in the largest component across institutes and DDC topics is shown in Fig. 3.10. When scholars are affiliated with many institutes and subjects (as advisees or advisors), they are counted as part of all the institutes and subjects to which they belong while determining the distribution of researchers by institutes and subjects.

According to Fig. 3.10a, a large number of researchers are affiliated with "Manonmaniam Sundaranar University" and "Alagappa University". "Vasudevan, T." [12](Alagappa University) is the researcher who has mentored the largest number of scholars in the main component of Shodhganga-AGN (see Fig. 3.9) and is partly responsible for the size of the largest connected genealogical subgraph in Shodhganga-AGN. In the largest genealogical subgraph, the majority of the researchers are from the science domain (see Fig. 3.10b). "Vasudevan, T." advised students in science (department of industrial chemistry). There are a few researchers in the largest genealogy graph who have worked in the technology area, and just three have worked in literature-related subjects/humanities.

[12]https://www.scopus.com/authid/detail.uri?authorId=7005045135

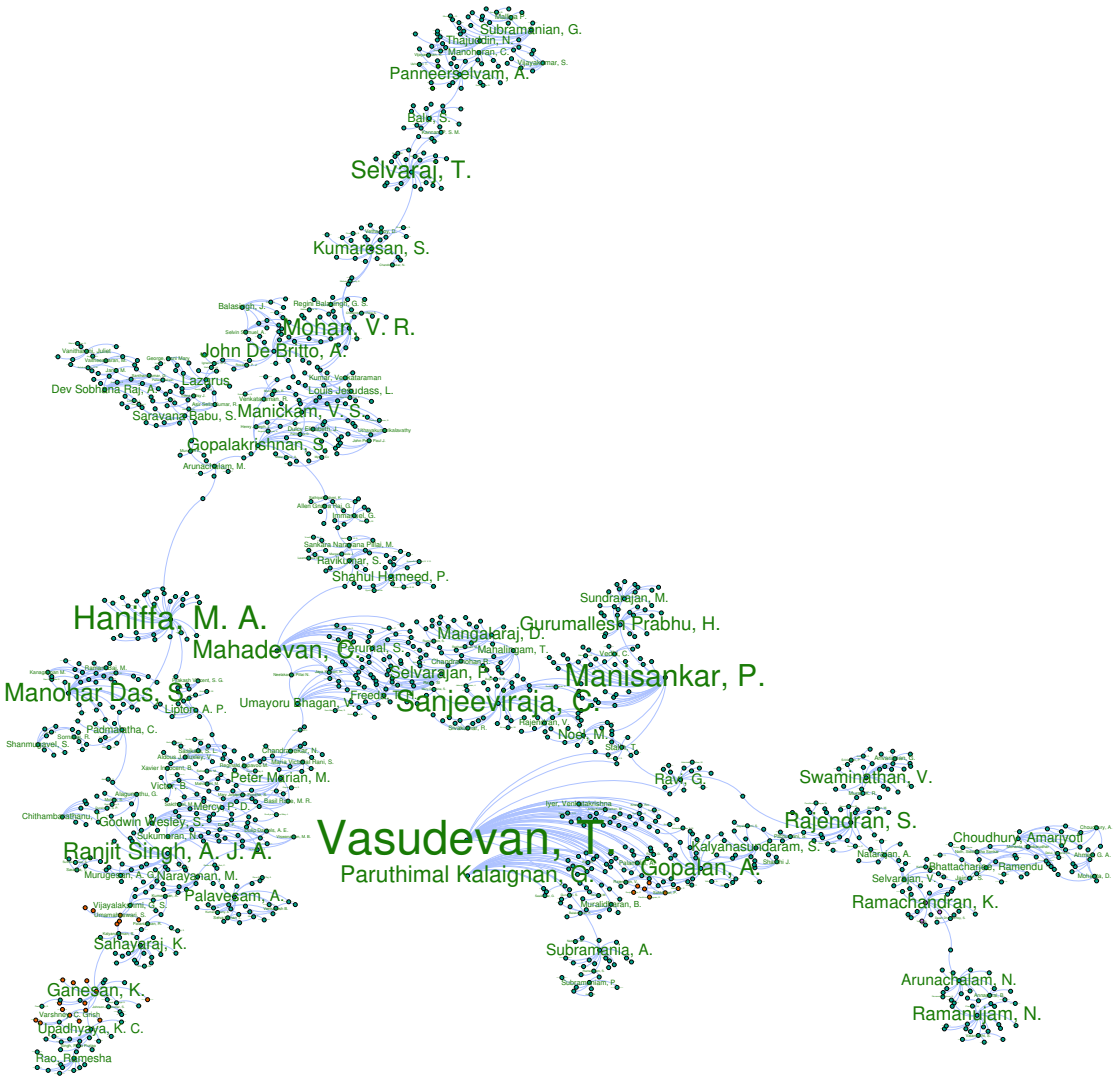
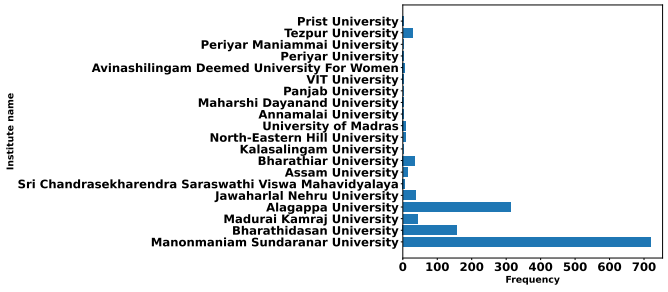
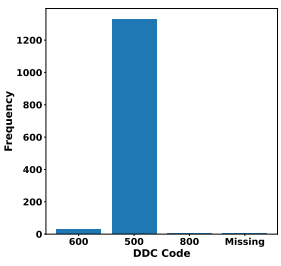


Figure 3.9: The largest connected component of the Shodhganga-AGN with researchers name. In the largest subgraph, the majority of the researchers are from the science domain, few researchers from technology domain, and three from literature-related subjects. The nodes/researchers are colored based on DDC subjects, node labels are proportional to fecundity value (larger node labels implies they have trained larger number of advisee). The Gephi force directed (ForceAtlas2) layout with property prevent overlap and expansion layout are used to generate the graph.



(a) Distribution of researchers across institutes.



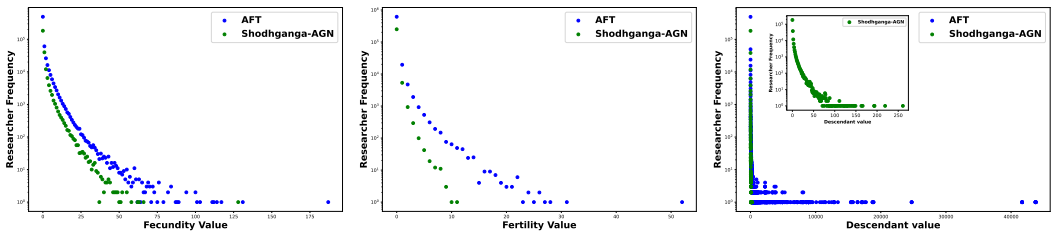
(b) Distribution of researchers across DDC subjects.

Figure 3.10: Distribution of researchers in the largest connected component of Shodhganga-AGN.

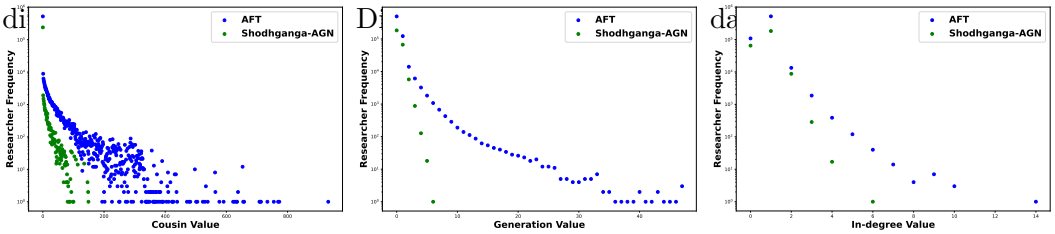
3.4 Analysis of Network Metrics and Subgraphs in AGN

3.4.1 Comparison of Metrics Distribution

In this section, we compared the metrics distribution of Shodhganga-AGN with the well-known database Academic Family Tree (AFT)[13]. The AFT is an interdisciplinary and heterogeneous database that maintains multiple relationships, which include graduate, research assistant, postdoc, collaboration, and post-graduate (Master, PhD). AFT is mainly crowd-sourced, which means it maintains information largely through voluntary user reports. In order to compare the metrics distribution, we only selected the relationship type "Post-graduate" which includes both master and PhD graduates from the AFT database. The filtered records constitute 643127 researchers and 554018 advisor-advisee relationships [14]. The number of weakly connected components in AFT and Shodhganga-AGN is 99889 and 56935, respectively. The distribution graphs are plotted in log-linear (semi-log plot) scaled mode.



(a) Comparison of Fecundity Distribution (b) Comparison of Fertility Distribution (c) Comparison of Descendant Distribution



(d) Comparison of Cousin Distribution (e) Comparison of Generation Distribution (f) Comparison of Multiplicity of Parents

Figure 3.11: Comparison of metrics distribution of Shodhganga-AGN and AFT

Fecundity [93, 89]: Fecundity is the number of academic children (direct descen-

[13]https://academictree.org/
[14]The May 2018 AFT data dump is used

dants) that a researcher has. In Fig. 3.11a, we can see that there is large number of researchers in Shodhganga-AGN and AFT for whom the fecundity value is low and very small number of researchers whose fecundity value is very high; overall, the distribution is positively skewed (heavy-tailed distribution) for both the genealogy network. We have listed the top five researchers from the Shodhganga-AGN, based on fecundity values, in Appendix Table 7.2.

Fertility [89, 93]: Fertility is the number of academic children who have a non-zero fecundity. We can see in Fig. 3.11b that there is a large number of researchers with very low fertility values, and very few researchers with very high fertility values. The fertility distribution follows power law for both the network. In Appendix Table 7.3, we have listed the top 5 researchers with the highest fertility in Shodhganga-AGN.

Descendants [89, 93]: Descendants are all advisees who have a direct or indirect mentoring relationship with the academician of interest. The distribution graph in Fig. 3.11c exhibits power law behavior (heavy-tailed distribution) for both Shodhganga-AGN and AFT, with a large number of researchers having a few descendants and very few researchers having a large number of descendants. We have listed the top 5 researchers from Shodhganga-AGN according to their descendant count in Appendix Table 7.4.

Cousins [89]: An academic who has a large number of cousins belongs to a large family and has prolific ancestors. The cousin distribution (Fig. 3.11d) for both networks begins with a high number of researchers with a low cousin count. At first, the distribution follows a power law decay trend. As the number of cousins increases along the x-axis, however, the trend changes to a zigzag pattern (ups and downs in the researchers' count as the cousin counts grows). We have listed the top five researchers from Shodhganga-AGN according to their cousin count in Appendix Table 7.5.

Generations [89, 93]: The number of generations of an academic is indicative of the impact, perpetuation, and evolution of their ideas and knowledge in the community to which they belong. The distribution graph for the metric in Fig. 3.11e shows similar behavior for both networks. We have listed the top five researchers

from Shodhganga-AGN according to the generation metric in Appendix Table 7.6.

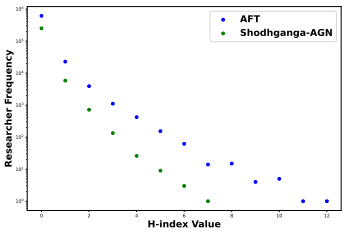
The Shodhganga-AGN comprises six generations of researchers, whereas the AFT involves more than 40 generations (Fig. 3.11e). The reason for the disparity is that Shodhganga-AGN started in 2011 and has yet to include the theses from other institutes in India, while AFT started in 2005. However, based on the available information, many of the institutes included in Shodhganga-AGN are younger than those included in AFT.

Multiplicity of Parents: We have reported the number of researchers who have multiple parents in Fig. 3.11f. The distribution graph in Fig. 3.11f shows that most of the researchers have 0 or 1 parent, and very few researchers have more than 3 parents/co-advisors in both networks. The distribution is positively skewed. In Appendix Table 7.7, we have listed the top 5 researchers that have the highest number of parents.

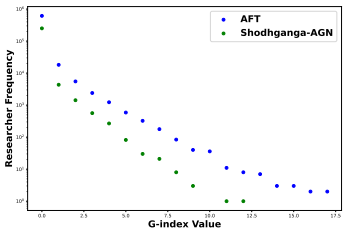
In general, the number of advisors who advise a student is 2 or 3, but in some cases there were 14 parents who advised a student in the graph, indicating that there is the possibility of noisy data in the academic genealogy graph.

Genealogy Indices [33, 94]: We calculated the h_m and g_m indices of Shodhganga-AGN and AFT researchers, as shown in Fig. 3.12a, 3.12b. Both h_m and g_m index distributions exhibit similar behavior for both networks. We have listed the top 5 researchers from Shodhganga-AGN based on the h_m and g_m index values in the Appendix Tables 7.8 and 7.9.

The Shodhganga-AGN metrics distribution was very similar to the AFT. That



(a) Comparison of h_m-index Distribution



(b) Comparison of G_m-index Distribution

Figure 3.12: Comparison of h_m-index and g_m-index distribution of Shodhganga-AGN and AFT

is, many researchers have low metrics values, while only a few have high values in both the dataset. The range of metrics values for Shodhganga-AGN differs

from AFT. First, this is because the number of researchers are less than the AFT. This may be due to incomplete researchers connections in Shodhganga-AGN and being country-specific, while AFT includes researchers from multiple nations. Also AFT started in 2005 and Shodhganga in 2011, which could also be possibly the reason why Shodhganga-AGN seems to have more missing connections than AFT, implying the smaller range values for metrics in Shodhganga-AGN than AFT. As of 2019, the Shodhganga repository also does not include theses from a number of India's premier national institutes of higher education and research.

3.4.2 Study of Identified Subgraph Structures in AGN

Shodhganga-AGN is a homogeneous AGN that captures only PhD advisor-advisee relationships extracted from the published thesis in the Shodhganga repository[15]. The network connects researchers from different fields. In order to understand the patterns of collaboration[16] among researchers in human resource training in higher education in India, we considered three distinct sub-graphs[17] (shown in Fig. 3.13a) based on our intuition and domain knowledge, to find out their occurrences within a large network defined over advisor-advisee relationships. We used a pre-built Neo4j[18] graph database pattern detection algorithm (Cypher Query Language, CQL) to obtain the frequency of occurrence of predefined subgraphs in Shodhganga-AGN.

We didn't consider temporal information while detecting sub-graph patterns from the genealogy network. In case of multiple advisors advising a researcher the thesis metadata do not include information about whether the advisors worked together to guide the researcher during his/her graduation period or at various intervals. Therefore, it is assumed that the advisers' duration of participation is equivalent to the researcher's period of graduation.

The counts of small subgraph patterns in graphs, are crucial for understanding

[15]We have made the assumption that the Researchers have only one doctoral degree

[16]The involvement of multiple advisors to advise a researcher in AGN

[17]Note: We have not identified the subgraphs based on their frequency of occurrence in the random graphs and Shodhganga-AGN.

[18]https://neo4j.com/

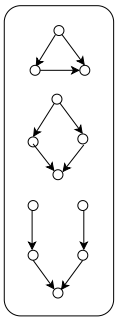
the structure of the complex network [83]. We investigate three distinct subgraphs to comprehend the supervision patterns[19] in the Shodhganga-AGN [111, 64]. Here, subgraphs of the Shodhganga-AGN captures one of the following patterns: (1) Researchers who have collaborated with their advisors to advise researchers (3-node subgraph in Fig. 3.13a), (2) Siblings (advised by the same advisor) who have collaborated to advise researchers (4-node subgraph in Fig. 3.13a), and (3) Advisors who have been advised by different advisors but came together to advise the researchers (5-node subgraph in Fig. 3.13a). In Fig. 3.13 we have shown the distribution for the identified subgraphs in Shodhganga-AGN. We can see from Fig. 3.13b that a large number of researchers who have been advised by different advisors have collaborated to advise other researchers (5-node motif). In most cases, researchers graduate from one institute and move on to another to share their knowledge and expertise.

There are only a few (28) researchers who have worked with their siblings (Fig. 3.13, 4-node subgraph). Researchers who graduated from the same institutes and were supervised by the same advisors may have moved to different institutes based on their personal preferences (institutes closer to their home town, city, or state; or leaving academia and moving into industry), and as a result, they have lost touch with their siblings. There are 224 researchers who have collaborated with their advisors to advise other researchers (Fig. 3.13, 3-node subgraph). In all such cases, researchers first graduated from the institutes and then became advisors at the same institutes. As a result, they had more opportunities to interact with their advisors and finally coordinated to advise the other researchers.

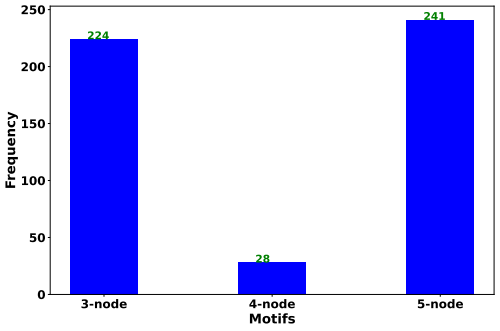
3.4.2.1 Distribution of subgraphs structure instances across institutes and subjects (DDC)

In this section, we present statistics on the institutions and DDC subjects prevalent in the 3-node, 4-node, and 5-node subgraph structures (shown in Fig. 3.13a) available within Shodhganga-AGN. All the extracted subgraph structures available within the Shodhganga-AGN are shown in Figs. 7.4a, 7.4b, and 7.4c

[19]Note: Temporal information is not considered for detecting motifs



(a) Subgraphs structure



(b) Subgraphs distribution

Figure 3.13: The 3-node, 4-node, and 5-node subgraphs and their distribution in Shodhganga-AGN

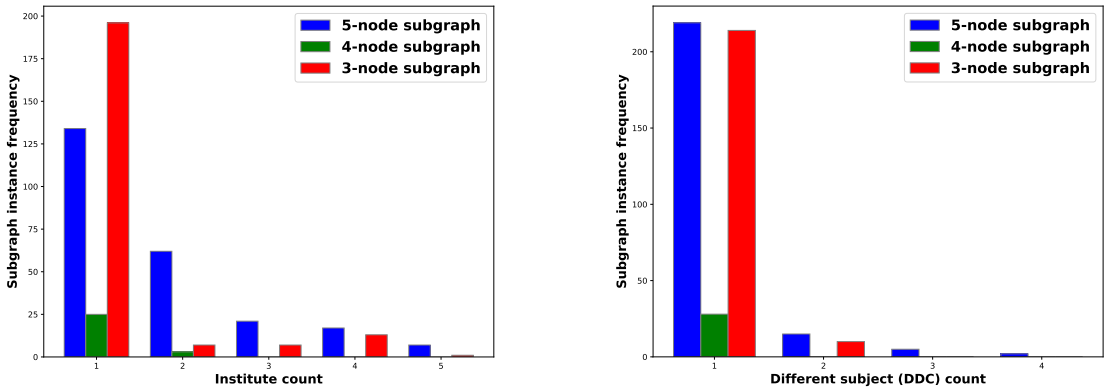
(Appendix Section 7.1.5). To calculate the number of distinct institutes for subgraph instances[20], we first obtained the institutes for all of the researchers in the subgraph instances (independent of the advisor-advisee relationship time, i.e., all affiliations for the researcher as a advisor or advisee are considered), and then calculated the number of distinct institutes. In the case of DDC subject codes, we first assign generic DDC subject codes to researchers (obtained based on their thesis and advised thesis i.e., if a researcher's theses DDC codes are 300, 310, 340, 400, 490 assigned with the generic DDC codes 300, 400.) and then count the number of unique DDC codes across subgraph instances.

Researchers have worked in multiple institutes, according to the distribution of institutes in Fig. 3.14a. We can see from Fig. 3.14a that a large number of subgraph structure instances belong to a single institute, implying that researchers were advisors and advisees in the same institutes, or that the details of the advisor affiliation (researcher as advisee or his/her advised theses information are missing) are missing in the datasets. Some researchers have worked in multiple institutes as advisors or advisees in the subgraphs under consideration, as shown in Fig. 3.14a.

The DDC subject count across the subgraph instances is shown in Fig. 3.14b. The researchers worked on up to four different subjects in the considered subgraphs. Most of the subgraph instances belong to a single domain, and a few belong to multiple domains, as shown in Fig. 3.14b. This implies that the vast majority of advisors have only trained advisees in a single subject/domain. There are some

[20]The advisor-advisee connections available in Shodhganga-AGN following the predefined subgraph structures are referred as instances of the subgraph structures (shown in Fig. 3.13a)

researchers who have worked in several different fields, as illustrated in Fig. 3.14b.



(a) Subgraphs instances distribution across institutes

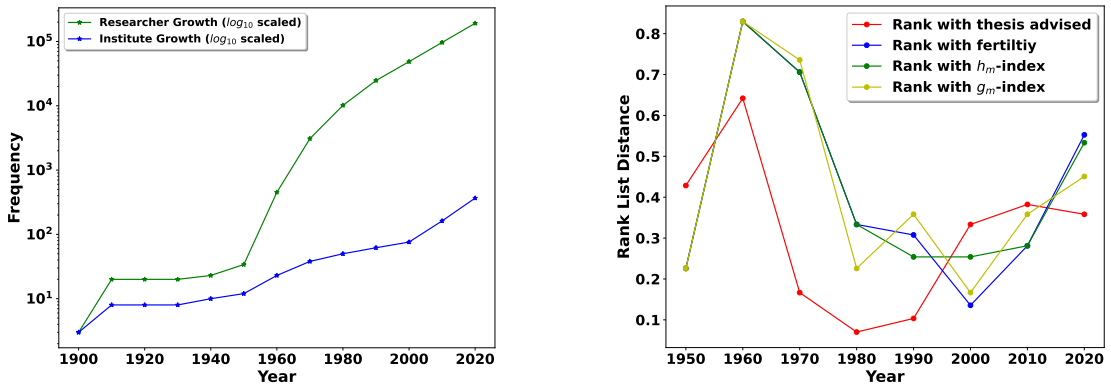
(b) Subgraphs instances distribution across subjects

Figure 3.14: Subgraphs instances distribution across institutes and subjects (DDC)

3.5 Role of Institutions

In this section, we look at genealogical metrics at the institute level in the Shodhganga-AGN. We focus on institute rankings, thesis distribution, and the increase in institutes over time in India. To obtain an institute's rank, we aggregated the metric values obtained by the researchers from that institute. The change in the ranks of the top ten institutes and growth in institutes over time is depicted in Fig. 3.15. To measure the change in order of ranks for the institute from the current year to the prior consecutive year, we used the modified Jaccard coefficient described in [41]. We also plotted a grid layout of the ranks in 20-year intervals in Fig. 3.16, with the color indicating the increase, decrease, or no change in the relative rank of two consecutively placed institute columns. The abbreviations of the institute/university names are provided in Appendix Table 7.11.

The distribution of researchers and universities/institutes over time is depicted in Fig. 3.15a. The distribution curves for researchers and institutes show a similar pattern, with relatively small increases between 1900 and 1950, followed by larger growth over time. In the case of institutions, we can see that the slope of the curve has changed significantly at a few stages, most notably in 1950, 1970, and 2000. The graph in Fig. 3.15a clearly shows that as the number of institutes expanded, so did the number of researchers who graduated. The cause might be the introduction



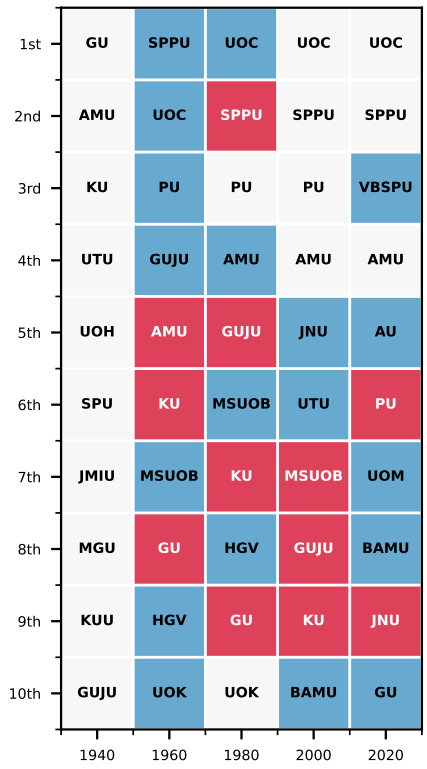
(a) Temporal evolution of universities/institutes and researchers

(b) Distance between institutes ranks over a 10-year interval with different metrics

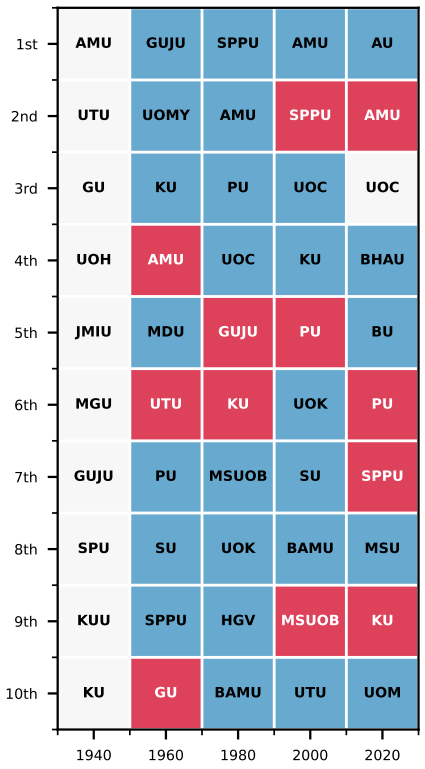
Figure 3.15: The growth of researchers/institutes over time and the distance between successive rankings of institutes

of Planning Commission following independence in 1950 and the appointment of the University Education Commission in 1948 for the improvement of the Indian education system.

The change in institute ranks at decade intervals is shown in Fig. 3.15b, and the area covered by institutes in terms of the number of researchers graduating is shown in Fig. 3.17. We calculated the distances between the top ten institutes' ranks over time intervals to study the temporal change in ranks. The change in institution rankings between 1980 and 2010, as illustrated in Fig. 3.15b, is slight, but the change in rank before 1980 and after 2010 is significant. Before 1950, only 39 researchers graduated from 12 universities, all of which were either public or central universities. After 1950, new public or central universities made it into the top 10 list, and the number of universities and graduates went up. This can be seen in Fig. 3.15a, which shows how the slope changed during 1950. The ranking of institutes changed the most during 1960 (see Fig. 3.15b), which could be because of the University Education Commission (also called the "Radhakrishnan Commission"), which was set up in 1948 as the first committee for higher education in independent India. Its job was to report on the state of Indian university education and make suggestions for improvements and expansions. The distance between the top 10 universities has decreased since 1960. Despite the fact that the institutes are expanding as depicted in Fig. 3.15a, this may be

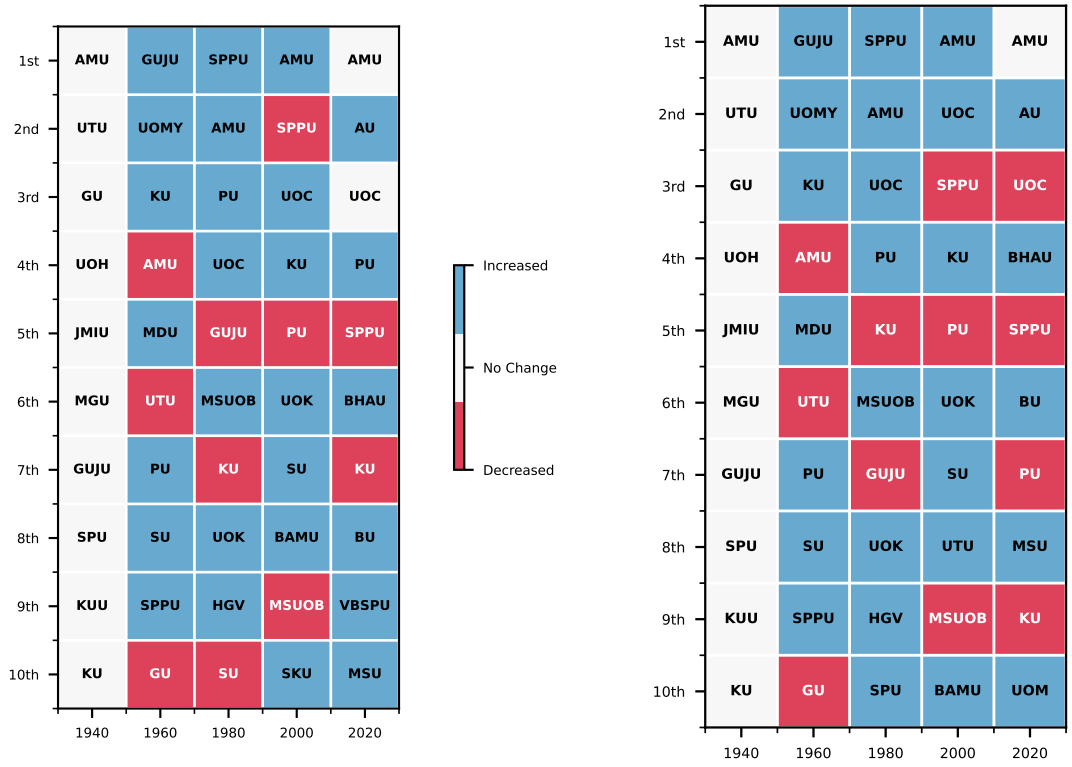


(a) Based on number of thesis advised



(b) Based on fertility

because the desired university (i.e., public university or central university) from which students were graduating was established. After India gained independence in 1947, this was also a time of growth for the country implying new institutes were growing and the institutes' ranks were changing, suggesting increase in the distance between the institute ranks over the time as shown in Fig. 3.15b. The shift in institute rankings between 2010 and 2020 is significant, but note that this could be a real tectonic shift or a consequence of missing data (advised thesis) for some institutes or in some periods of time. The graphs in Fig. 3.15b are very correlated for the metrics g_m-index, h_m-index, and fertility, while for advised theses it is slightly different. The reason is former metrics focus on advisees of advisors who become advisors themselves, while later only consider advised students. There are 28 universities/institutes that have appeared over the years among the top 10 institutes based on different metrics (refer Appendix Table 7.11) in the Shodhganga-AGN. The universities/institutions that once appeared in the top ten and then continued to appear in subsequent time intervals based on



(c) Based on g_m-index

(d) Based on h_m-index

Figure 3.16: Institute rank plotted over 20 year intervals with different metrics (Institute rank is calculated based on aggregation of researchers or researchers metrics (check Appendix Table 7.11 for institute abbreviation))

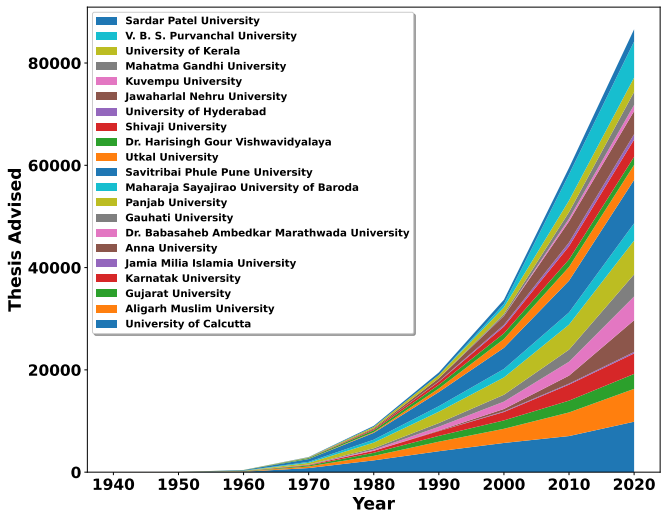


Figure 3.17: Institutes area plot based on researchers graduating over time

metrics, thesis published, fertility, g_m, and h_m are the University of Calcutta (CU), Aligarh Muslim University (AMU), Panjab University (PU), Karnatak University (KU), Anna University (AU). These institutes, except Karnatak University (rank

not provided), have also appeared in the top 30 universities based on the NIRF University (National Institutional Ranking Framework, 2021) ranking[21], released by the Ministry of Education, Government of India. The institutes/universities appearing at the top of the NIRF rating did *not* appear in our list because most of them are IITs / NITs / IISc, that do not submit their theses to the Shodhganga repository. Moreover, the parameters for NIRF or any other ranking framework are typically much broader than the set of AGN-based metrics that we consider here.

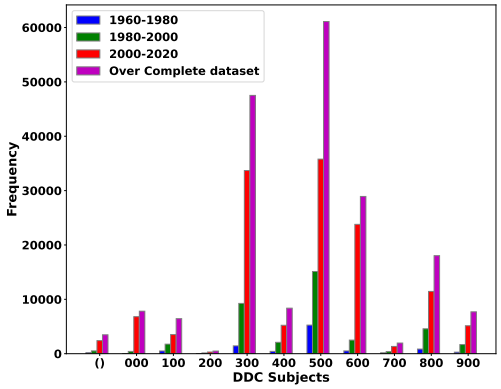
3.6 Subject Domain Analysis

In this section, we look at the aggregation of genealogical metrics at the subject level. We focus on the temporal distribution of researchers across subjects, the areas covered by subjects in terms of graduating researchers, and the distance [41] between the subjects' ranks (obtained by aggregating the metrics value of researchers affiliated with subjects) over a decade interval. To study the temporal growth of the subjects, we calculated the aggregate of the researchers' metrics across the subjects over subsequent time intervals. Note that if a researcher's thesis pertains to many subjects, they are considered to be a part of all of those subjects while conducting the aggregate. Fig. 3.18a shows the growth of the subjects in terms of the number of students who graduated in these subjects. According to Fig. 3.18a, the most popular subjects from 1960 to 1980 were Science, Social Science, Literature, Language, Philosophy and Psychology, History and Geography, and Technology, in that order. Computer Science, and Arts and Recreation emerged in subsequent decades.

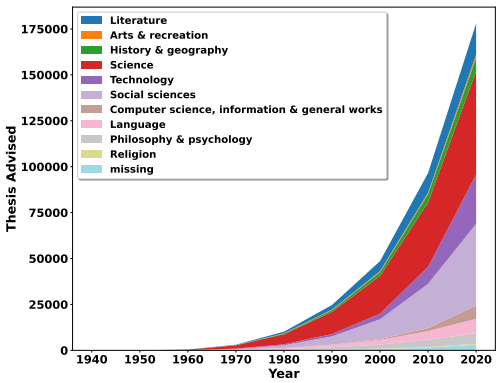
The area covered by the subjects over time is depicted in Fig. 3.18b. According to Fig. 3.18b, a large proportion of the total area is covered by only two subjects, namely Science and Social Science. This implies that most of the researchers have done their research on topics related to Science, and Social Science.

The distance between the ranks of the top ten subjects over a ten-year period

[21]https://www.nirfindia.org/2021/Ranking.html



(a) Researchers distribution across DDC subjects over time



(b) DDC subject area plot in terms of researchers graduating over time.

Figure 3.18: Temporal evolution of researchers' across DDC subjects. The order of the subject plot are same as the legends in Fig. (b), i.e. first plot refers to "Literature" and last refers to "missing" (() or word "missing" denotes DDC subject codes for the published theses are missing; check Appendix Table 7.10 for DDC code names).

is depicted in Fig. 3.19. Initially, the number of domains in which students were graduating was growing (emergence of new domains). Therefore, the changes in the ranks were initially high. But after 1980, the ranks of the top subjects did not change very drastically; the reason could be the saturation of domains (in other words, most of the domains in which researchers are graduating have emerged) in which researchers were graduating.

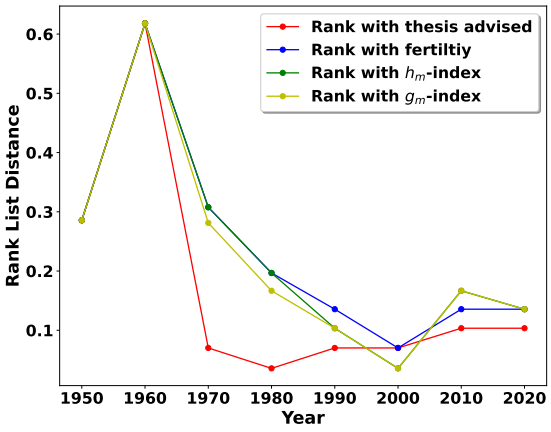


Figure 3.19: The distance between consecutive DDC-based subject ranks with different metrics over 10 year intervals (The DDC subject ranks are calculated by aggregating the metric values of the researchers from the subjects at a given time).

3.7 Chapter Summary

We conducted a variety of analyses in the Shodhganga Indian Thesis Repository. Firstly, we disambiguated the researchers' names and then generated an academic genealogy graph called Shodhganga-AGN. We analyze the AGN with the help of various genealogical metrics. We also examined the significance and contribution of universities and subjects, as well as the collaboration patterns of researchers in human resource training in the Indian Education System. The researchers' metrics have been able to determine the universities and academics who have significantly contributed to the growth of India's higher education system. There are 1356 researchers and 1437 advisor-advisee relationships in the largest connected component of Shodhganga-AGN. The component primarily consists of scientists, who are mainly connected to three institutions. We examined subgraphs in the genealogical network to find patterns of supervision and discovered that the majority of subgraph instances link researchers to the same institution or field of study. A limitation of the present work is that it considers only Shodhganga as the data source. However, many premier national institutions of higher education and research in India, such as the Indian Institute of Science, Indian Institutes of Technology, and the Indian Statistical Institute, maintain their own ETD repositories that are not present in Shodhganga currently. So one of our future goals is to consider these repositories for AGN analysis.