



“Analyzing the social influence on the evolution of scientific citations and collaborations”

- Project Report

ABSTRACT

Recently, conference publications have gained a wide popularity, especially in the domain of computer science. In conferences, the opportunity of personal interactions between the fellow researchers opens up a new dimension for the citation network evolution. On the other hand, it is also very important to study the development of scientific collaborations among the researchers. In this work, we propose a generic multiplex network framework to uncover the influence of the interactions in a conference on the appearance of the new citation links as well as to systematically study the impact of past citation links on the creation of future collaborations. We crawl the DBLP citation dataset and perform a case study on the leading conferences in the “Networking and Distributed Systems”, “Artificial Intelligence”, “Hardware & Architecture” and “Human-Computer Interaction” domains. Our experiments are able to identify the "successful" conference interactions and "successful" citations, which eventually results in the induced citations and the induced collaborations respectively. Interestingly, it is also found that a fast influence mostly results into a sustained influence.

The shortlisted conferences domain wise are as follows:

Artificial Intelligence

Topic:

Artificial Intelligence	(#artificial_intelligence)
Information Retrieval	(#information_retrieval)
Natural Language NLP and Speech	(#fnatural_language_and_speech)
Machine Learning and Pattern Recognition	(#fmachine_learning_and_pattern_recognition)
Data Mining	(#fdata_mining)
Bioinformatics & Computational Biology	(#fbioinformatics_and_computational_biology)

	dblp link	session_info	original count	topic count
ICRA	http://www.informatik.uni-trier.de/~ley/db/conf/icra/	1993-2012	8639	8638
NIPS	http://www.informatik.uni-trier.de/~ley/db/conf/nips/	1988-2012	3598	3587
AAAI	http://www.informatik.uni-trier.de/~ley/db/conf/aaai/	1980-2013	3375	3367
ICDE	http://www.informatik.uni-trier.de/~ley/db/conf/icde/	1984-2013	2666	2616
SIGIR	http://www.informatik.uni-trier.de/~ley/db/conf/sigir/	1979-2013	2074	1941
ECAI	http://www.informatik.uni-trier.de/~ley/db/conf/ecai/	1984-2012 (2yr gap)	2110	2096

Networking and Distributed Systems

Topic:

Networking	(#fnetworks_and_communications)
Distributed and Parallel Computing	(#fdistributed_and_parallel_computing)
World Wide Web	(#fworld_wide_web)

	dblp link	session_info	original count	topic count
ICPP	http://www.informatik.uni-trier.de/~ley/db/conf/icpp/	1983-2012	1299	1297
INFOCOM	http://www.informatik.uni-trier.de/~ley/db/conf/infocom/	1989-2013	4272	4270
IPDPS	http://www.informatik.uni-trier.de/~ley/db/conf/ipps/	1992-2013	3465	3449
ICDCS	http://www.informatik.uni-trier.de/~ley/db/conf/icdcs/	1982-2012	1968	1964
WWW	http://www.informatik.uni-trier.de/~%20LEY/db/conf/www/	1995-2013	1262	1219
ICC	http://www.informatik.uni-trier.de/~ley/db/conf/icc/	2000-2012	2551	2534
GLOBECOM	http://www.informatik.uni-trier.de/~ley/db/conf/globecom/	2000-2012	3105	3089

Hardware & Architecture

Topic:

Hardware & Architecture
Security and Privacy
Operating Systems
Real Time Embedded Systems

(#fhardware_and_architecture)
(#fsecurity_and_privacy)
(#foperating_systems)
(#freal_time_and_embedded_systems)

	dblp link	session_info	original count	topic count
ISCAS less	http://www.informatik.uni-trier.de/~ley/db/conf/iscas/	1993-2013	5387	5384
DAC	http://www.informatik.uni-trier.de/~ley/db/conf/dac/	1990-2013	3805	3805
DATE	http://www.informatik.uni-trier.de/~ley/db/conf/date/	1994-2013	2625	2625
ICCD	http://www.informatik.uni-trier.de/~ley/db/conf/iccd/	1991-2012	1712	1711
ICCAD less	http://www.informatik.uni-trier.de/~ley/db/conf/iccad/	1990-2012	2312	2312
CRYPTO	http://www.informatik.uni-trier.de/~ley/db/conf/crypto/	1984-2013	1063	1049

Human-Computer Interaction

Topic:

Human-Computer Interaction
Multimedia
Graphics
Computer Vision

(#fhuman-computer_interaction)
(#fmultimedia)
(#fgraphics)
(#fcomputer_vision)

	dblp link	session_info	original count	topic count
ICIP	http://www.informatik.uni-trier.de/~ley/db/conf/icip/	1993-2012	3611	3441
ACM MULTIMEDIA	http://www.informatik.uni-trier.de/~ley/db/conf/mm/	1994-2012	1824	1822
CVPR	http://www.informatik.uni-trier.de/~ley/db/conf/cvpr/	1996-2012	1918	1918
MVA	http://www.informatik.uni-trier.de/~ley/db/conf/mva/	1988-2011 (2yr gap)	1244	1243

(domain_name stands for art_int, hw_archi, nw_dstrbd, human_comp)

- Using the DBLP links of the above mentioned conferences data was collected for the respective years and kept in **domain_name_conference_sessions** under **conferencename_year** text files. (Note: Workshops were not taken into account)

Screenshot:



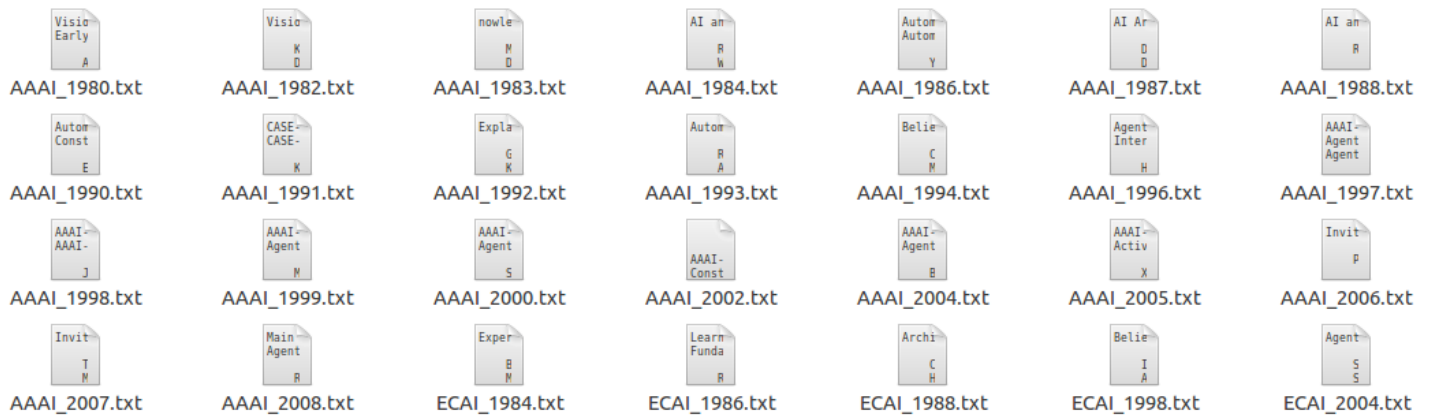
The DBLP Computer Science Bibliography

art_int_conference_sessions

human_comp_conf_sessions

hw_archi_conf_sessions

network_dstrbd_conference_sessions



- After gathering the data for all the years for the shortlisted conferences for all domains, the program **flexi_session_layer.c** was executed on each domain's conference sessions to get all possible interaction combinations for the following cases for each domain
 - If only first author was present
 - If only last author was present
 - If there was a probability that either first or last or both the authors go – (3 cases)
 - 0.45_0.45_0.1 (first_last_both)
 - 0.40_0.40_0.2 (first_last_both)
 - 0.35_0.35_0.3 (first_last_both)

flexi_artint_session_files

flexi_humancomp_session_files

flexi_hwarchi_session_files

flexi_nwdstrbd_session_files

flexi_session_layer.c

flexi_session_first_author

flexi_session_last_author

flexi_session_mixed_0.4_0.4_0.2

flexi_session_mixed_0.35_0.35_0.3

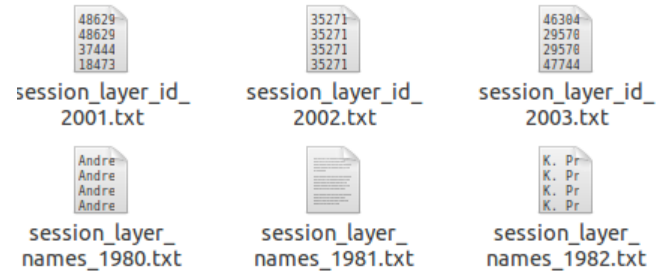
flexi_session_mixed_0.45_0.45_0.1

- The above program takes conference data from domain_name_conference_sessions for all the years as input along with **citations.txt**, **unique_authors.txt**, **year_conferences.txt** and **unique_authors_id.txt** and gives all possible interaction combinations between the authors in their respective sessions for each year

```

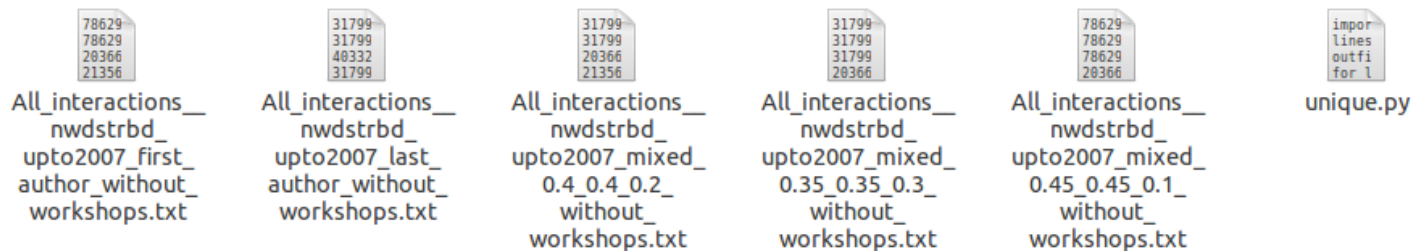
78629 20366 ICDCS 1982
78629 77883 ICDCS 1982
20366 77883 ICDCS 1982
213568 109240 ICDCS 1982
213568 18895 ICDCS 1982
109240 18895 ICDCS 1982

```

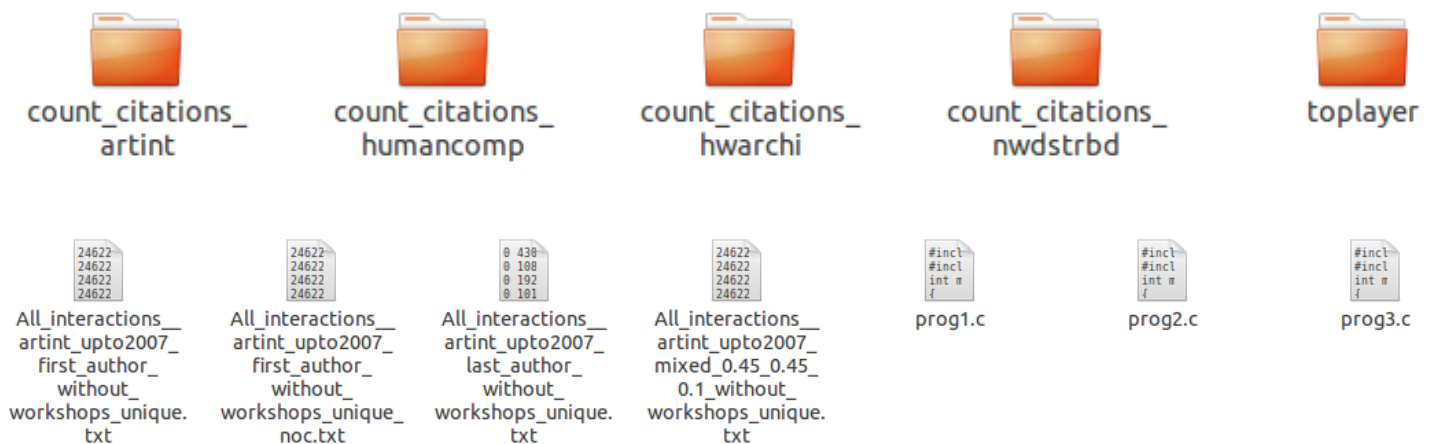


(flexi_domainname_session_files/diff_cases – contains year wise author names as well as author id's along with the conference names where they might have interacted in two separate text files) as well as a single txt file for all years combined which is stored in domain_name_links/nonunique.

- Therefore we get five files for each domain (first, last, 3 probabilities).
- We use **unique.py** over each file for 20 files (5 files X 4 domains) to remove the duplicate entries which is stored in domain_name_links/unique.



- Next what I did was to find the number of successful interactions for first author, last author, author with probability 0.45_0.45_0.1 files for all domains using **prog1.c**, **prog2.c** and **prog3.c** (I got the citation data from toplayer) (Note: The output files interaction_noc.txt are present on the server.)



- An interaction is said to be successful if it leads to a citation in future years.

- After finding out the number of successful interactions for all the 3 files for all domains we divide them with the 3 total interactions file respectively to get the domain wise conversion rates which are as follows:

➤ AI Domain

- ✓ first_author - 3.1% (Interactions 81023, Successful 2544)
- ✓ last_author - 4% (Interactions 73724, Successful 2958)
- ✓ author_0.45_0.45_01 - 3.7% (Interactions 94570, Successful 3539)

➤ Humancomp Domain

- ✓ first_author - 0.6% (Interactions 103867, Successful 675)
- ✓ last_author - 3.31% (Interactions 93643, Successful 3108)
- ✓ author_0.45_0.45_01 - 1.68% (Interactions 124563, Successful 2101)

➤ Hwarchi Domain

- ✓ first_author - 6.9% (Interactions 15160, Successful 1060)
- ✓ last_author - 12.6% (Interactions 14571, Successful 1836)
- ✓ author_0.45_0.45_01 - 9.8% (Interactions 17651, Successful 1744)

➤ Nwdstrbd Domain

- ✓ first_author - 1% (Interactions 50514, Successful 531)
- ✓ last_author - 2.1% (Interactions 46516, Successful 1023)
- ✓ author_0.45_0.45_01 - 1.7% (Interactions 57743, Successful 977)

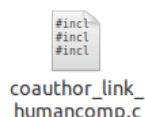
- After this I found out all combinations of coauthor links for all domains using the program coauthor_link_domainname.c which takes **citations.txt**, **unique_authors.txt** and **unique_authors_id.txt** as input and produces year wise coauthor link text files in **domain_name_coauthor** directory. Each file has the format Author1_id Contname1 Author2_id Contname2 Paper_Index Conf_name Year.



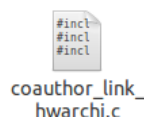
citations.txt



coauthor_link_
artint.c



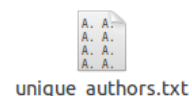
coauthor_link_
humancomp.c



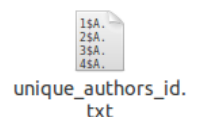
coauthor_link_
hwarchi.c



coauthor_link_
nwdstrbd.c



unique_authors.txt



unique_authors_id.
txt



- Following the generation of coauthor links I filtered the above year wise output files based on the fact that each coauthor participating in the coauthor link must have at least one previous publication before the co-authorship year and also removed duplicate entries (if present) using the program **filter_publish.c** which takes **citations.txt**, **unique_authors.txt** and **unique_authors_id.txt** as input and produces year wise coauthor link text files in **domain_name_coauthor_refined** directory along with **first_publish.txt** which consists of author id and his first publication year for all the 501060 authors. Each file has the format Author1_id Contname1 Author2_id Contname2 Paper_Index Conf_name Year.

```
Author1 @321034 #2 Author2 @377114 #4 paper &11695 conf *AAAI year !1980
Author1 @321034 #2 Author2 @385569 #2 paper &11695 conf *AAAI year !1980
Author1 @321034 #2 Author2 @322535 #2 paper &11695 conf *AAAI year !1980
Author1 @377114 #4 Author2 @385569 #2 paper &11695 conf *AAAI year !1980
```



- It was also proposed that after generation of coauthor links for all the domains intersection between same authors publishing in multiple domains can be found.
- The following sources were identified as factors for co-authorship:-
 - Previous citations
 - Previous interactions
 - Attended same conference in same year
 - Previous co-authorship
 - In fields of same domain
 - In mixed fields of different domains

- An account was created on “Mendeley” website so that personal information about authors could be crawled for and incorporated within the project but after signing up which required name, email, password, field of study (arts, computer, chemical), academic status (Btech, Phd, Researcher, Student) it was found that no such information about popular authors could be found. Other less noteworthy authors also uploaded their achievements, awards, publications, qualification, etc. but no personal information. It was basically a platform for individuals to upload their papers and ask for suggestions or clear their queries with other members of the same or similar field by joining a group and interacting with them or by adding them as friends by sending a request. The website also provided a desktop software for uploading documents and to manage and keep them arranged on the internet.



Create free account

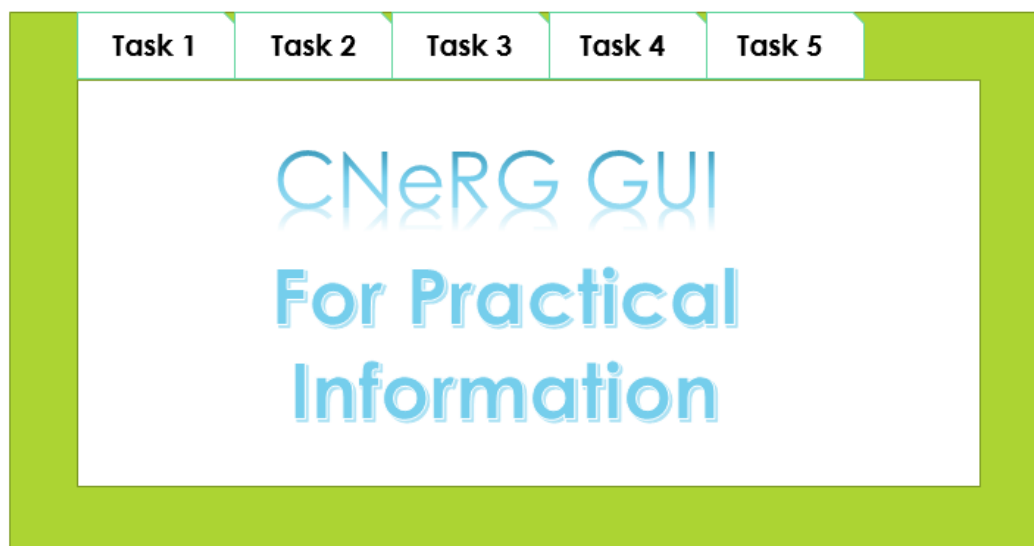
Sign in

Join millions of researchers today

Mendeley streamlines your workflow, saving you time to focus
on what is important.

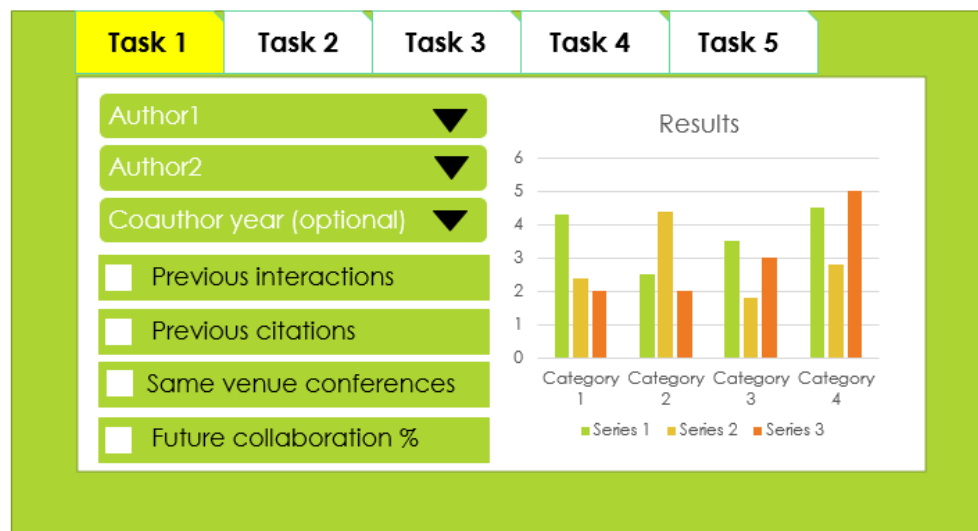
- Next we planned to design a GUI which could perform a set of tasks which the user cannot find on the internet and only then the whole project would fulfill some useful purpose. The design and tasks that can be performed by the GUI are as follows:

FUNCTIONS OF THE USER INTERFACE



Task 1 – Other Author related info

- ▶ Inputs – Author1, Author2, Coauthored Year (optional)
- ▶ Provide info about :-
 - ▶ Previous interactions between them (can also mention conversion rate between them)
 - ▶ Previous citations between them (lifespan of induced citation between them)
 - ▶ Number of same venue conferences attended
 - ▶ Future collaboration possibility (considering mediators between the above coauthors, same or different continents they belong to, other facts)
- ▶ Output can be displayed as top 10 results in the first page and then can be browsed to 2nd and so on depending on the results.



Task 2 – Field Migration Recommendations


- ▶ Inputs – Author name, field of study, domain
- ▶ Output – Give recommendations to switch field based on the following two parameter
 - ▶ Popular authors who have publications in multiple fields/domains
 - ▶ Coauthors of the above mentioned author who have published in multiple fields/domains
- ▶ Suggest a list of conferences that the author should submit his research papers.

Task 1Task 2Task 3Task 4Task 5

Author name
Field of study
Domain

☐ Popular authors who have published in multiple fields/domains
☐ Coauthors who have published in multiple fields/domains

Fields



1st Qtr
2nd Qtr
3rd Qtr
4th Qtr

Task 3 – Collaboration with specific author

- Input Author1 (amateur author), Author2 (the popular author he wants to coauthor with)

The following steps should be followed if author1 wants to collaborate with author2

- The no. of papers author1 should cite should be - ____
- The no. of interactions author1 should indulge in with author2 should be - ____
- The no. of same venue conferences author1 should attend - ____

If the above steps are followed then there is __% chance of collaboration with the specified author after __ years.

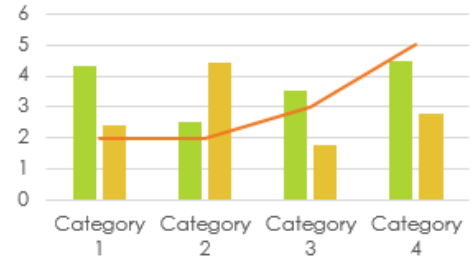
(Note: The blank spaces can be calculated by taking into account publications of both the authors and other statistical information available.)

Task 1Task 2Task 3Task 4Task 5

Author1 (Amateur)
Author2 (Big shot)

Author 1 should :-
☐ Cite this much no. of papers
☐ Have this much no. of interactions
☐ Attend this much no. of same venue conferences

Recommendations & Suggestions

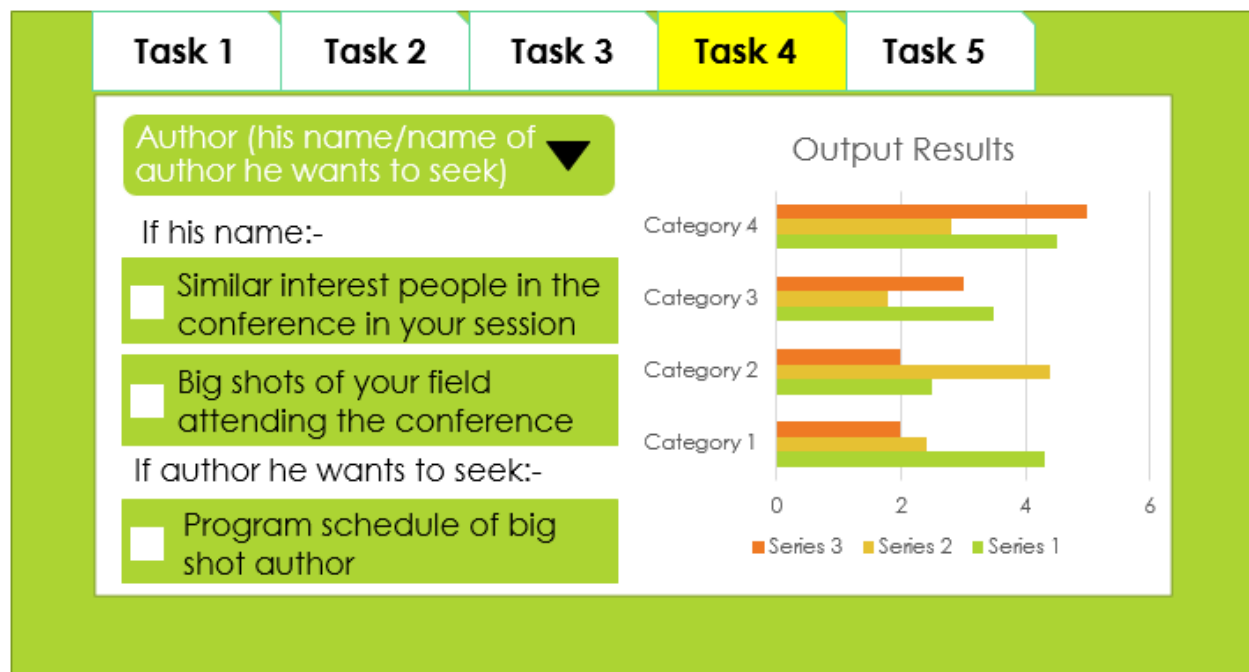


Category	Series 1	Series 2	Series 3
Category 1	4.2	2.2	2.0
Category 2	2.5	4.5	2.0
Category 3	3.5	1.8	3.0
Category 4	4.5	2.8	5.0

Series 1
Series 2
Series 3

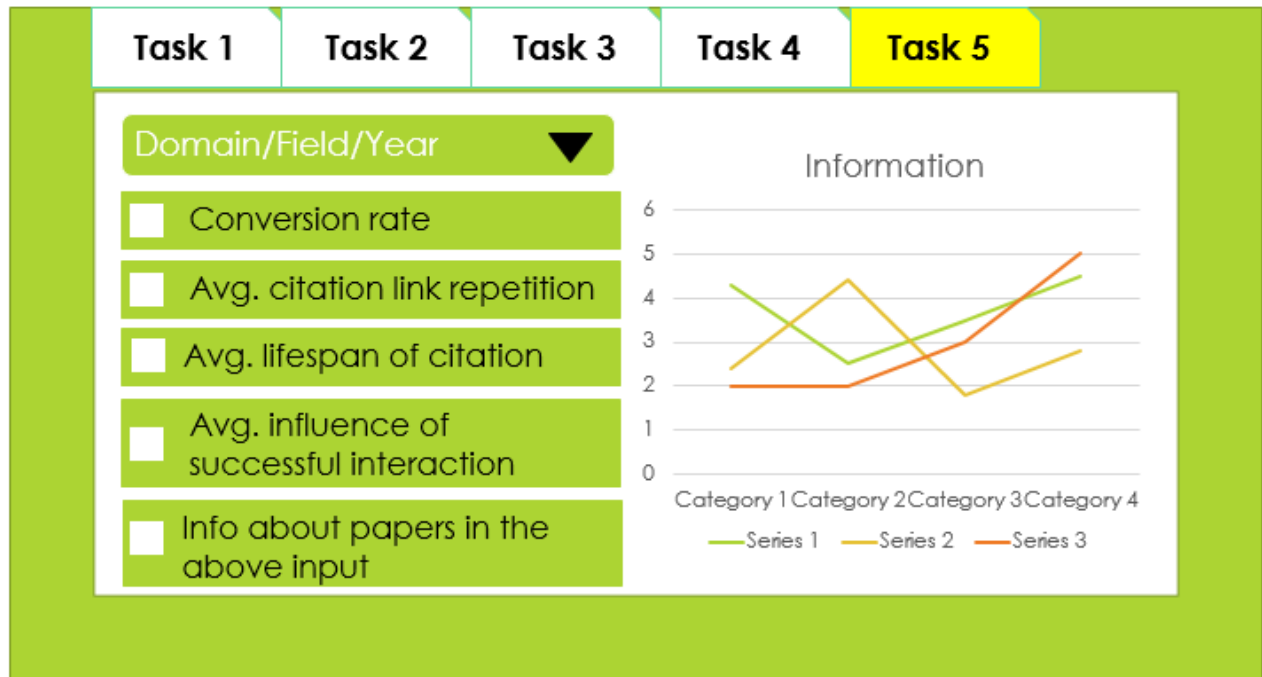
Task 4 – Program Schedule

- ▶ **Input – Authors name or author who wants to seek**
- ▶ **Provide info about –**
 - ▶ Program schedule i.e. the conference name, date, etc. where the author he wants to seek would be going
- If user enters his/her name then
 - ▶ Similar interest people in the conference in your session
 - ▶ Big shots of your field attending the conference
- ▶ (Note: In order to provide the above info the UI must be kept updated with the latest schedules for the upcoming conferences.)



Task 5 – Domain wise Information

- ▶ **Input – Domain**
- ▶ **Provide info about –**
 - ▶ What is the conversion rate(successful interactions in all conf/total interactions in all conf) in that domain
 - ▶ What is the average induced citation link repetition
 - ▶ What is the average lifespan of induced citation
 - ▶ What is the average influence of successful interaction (diff between successful interaction & 1st citation)
 - ▶ What is the % of successful interaction if authors belong to same/diff continents
 - ▶ All papers published in the above domain which can be filtered by author name/conference/year
- ▶ (Note: The above info can also be provided if instead of domain the input is particular field/year)



- After this I listed down the features that would influence the **Task 1**

1. Interaction count using 0.45_0.45_0.1 author file
2. Citation count
 - a. Unidirectional
 - b. Bidirectional
3. Common Co-author count
4. Conversion rate/ Average influence gap
5. Same conference same year
6. Same/Different continent
7. Number of common fields
8. Co-authorship count
9. Year gap between last interaction and current year
10. Year gap between last citation and current year
11. Year gap between last co-authorship and current year
12. Standard deviation of citation years
13. Standard deviation of co-authorship years
14. Citation count difference
15. Publication count difference

- Next using **select_author.c** I selected 13000 random author pairs by taking 500 author pairs from each year from 1982 to 2008 except 1995 using citation files from top layer. The random authors generated were stored in **random_author_pair_4.txt** with the format Author1_id Author2_id year.











```

17233 243603 1982
337549 204068 1982
318098 142925 1982
153811 121479 1982
433160 185820 1982

```

random_author_pair_4.txt **select_author.c**

- Then using **random_author_pair_4.txt**, 1-4.txt (0.45_0.45_0.1 author file of all 4 domains) as input the program **count_interactn_1.c** produced files **interaction_noc_unique_new.txt** which gives number of interactions between author pairs and the year in which the interaction occurred and another file **interaction_noc_unique_finite.txt** which only gives interactions between authors pairs which is more than 0 followed by year in which it occurred followed by difference from the year 1995.








1.txt **2.txt** **3.txt** **4.txt** **count_interactn_1.c** **interaction_noc_unique_finite.txt** **interaction_noc_unique_new.txt**

interaction_noc_unique_new.txt	interaction_noc_unique_finite.txt
1 @17233 @243603 of year *1982 interacted #0 times	1 @223745 @98124 of year *1983 interacted #1 times !1982 - *13
2 @337549 @204068 of year *1982 interacted #0 times	2 @470566 @399781 of year *1984 interacted #1 times !1984 - *11
3 @318098 @142925 of year *1982 interacted #0 times	3 @344417 @500216 of year *1985 interacted #1 times !1987 - *8
4 @153811 @121479 of year *1982 interacted #0 times	4 @344768 @456316 of year *1985 interacted #1 times !1988 - *7
5 @433160 @185820 of year *1982 interacted #0 times	5 @444843 @140080 of year *1986 interacted #1 times !1982 - *13

- After that using **prog3.c** and **random_author_pair_4.txt**, toplayer as input it produces **citations_noc_new.txt** which gives number of times author pairs have cited one another and vice versa along with the years and if citations are before or after the year 1995.

```

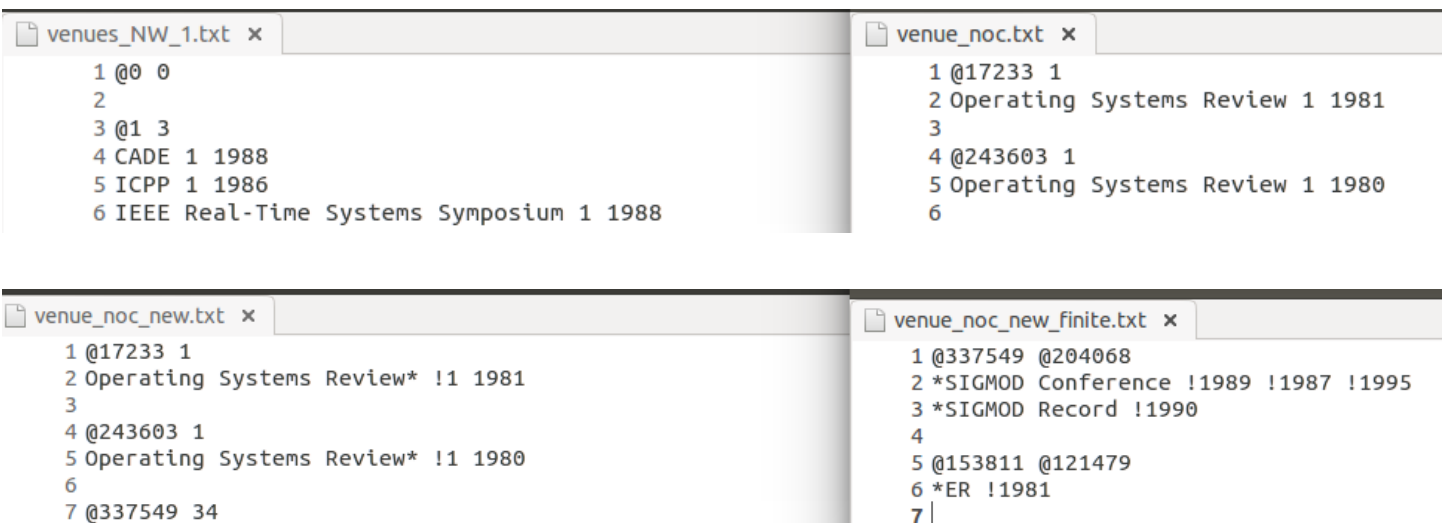
@17233 @243603 of year *1982 cited #1 <1995 #0 >1995 !1982
@243603 @17233 of year *1982 cited #0 <1995 #0 >1995
@337549 @204068 of year *1982 cited #1 <1995 #0 >1995 !1982
@204068 @337549 of year *1982 cited #0 <1995 #0 >1995
@318098 @142925 of year *1982 cited #1 <1995 #0 >1995 !1982
@142925 @318098 of year *1982 cited #0 <1995 #0 >1995
@153811 @121479 of year *1982 cited #1 <1995 #0 >1995 !1982
@121479 @153811 of year *1982 cited #0 <1995 #0 >1995
@433160 @185820 of year *1982 cited #1 <1995 #0 >1995 !1982
@185820 @433160 of year *1982 cited #0 <1995 #0 >1995
@347635 @207289 of year *1982 cited #1 <1995 #0 >1995 !1982
@207289 @347635 of year *1982 cited #0 <1995 #0 >1995
@187288 @93784 of year *1982 cited #3 <1995 #0 >1995 !1982
@93784 @187288 of year *1982 cited #0 <1995 #0 >1995

```


- Later, using the program **same_continent.c** and taking **random_author_pair_4.txt**, toplayer as input the output file **continent_noc_new.txt** was generated which had the format Author1_id Cont1_id Author2_id Cont2_id and determined whether both the authors belonged to same continent or not.

```
@327016 #4 @118033 #2
@390990 #0 @1862 #2
@21571 #2 @61131 #2
@243205 #4 @115302 #2
@399257 #0 @302064 #2
Total authors with same continent 7001 out of 13000
```

- For same venue same conference the following things were done. Firstly using **same_venue.c**, **random_author_pair_4.txt** and **VENUES_NW_1.txt** (it contains author id wise conferences attended and the years in which it was attended for all 501060 authors) as input, **venue_noc.txt** was created which contained information about only the 13000 random author pairs. Secondly using **venue_noc.txt** as input the program **same_venue_2.c** was run so as to insert proper symbols so that the file could be read producing **venue_noc_new.txt**. Using the previous output file and program **same_venue_1.c** the file **venue_noc_new_finite.txt** was generated which contained info about random author pairs and the same conference names and same years on which they attended it. The previous program also produces **noc_publish.txt** which consists of author id and their respective number of total publications for each author pair having the format Author1_id Publications1 Author1_id Publications2.



noc_publish.txt x

```
1 @17233 pub #1 @243603 pub #1
2 @337549 pub #67 @204068 pub #67
3 @318098 pub #1 @142925 pub #5
4 @153811 pub #3 @121479 pub #74
5 @433160 pub #7 @185820 pub #4
6 @347635 pub #45 @207289 pub #1
```

- Programs `coauthor_count.c` and `same_field.c` were also written and executed on server taking `citations.txt`, `unique_authors.txt`, and `unique_authors_id.txt` as input and produces `common_coauthor.txt` and `same_field.txt` having output formats respectively as

@Author_id no_coauthors

*cauthor_1 \$no_years !year1 !year2

*coauthor_2 \$no_years !year1

@Author_id

*field_name1 !year1 !year2

*field_name2 !year1

- After collecting all the outputs that influence Task1 these features have to be given as different test cases to SVM in order to implement machine learning along with certain class 0 or 1. After the training is complete for say 12000 random authors out of 13000 the remaining 1000 can be used for prediction by machine learning and its accuracy has to be checked. Thus by implementing the above the GUI can be used by any user.
- Then I listed down the features that would influence the **Task2**
 1. 23 fields -> 5/6/7 domains
 2. 18771 conferences
 3. Domain – Conf1 field1 field2 field3
 4. Author wise year wise publications in different domains
 5. Similarly do the above for big shots in the same field as author input
 6. Calculate migration index (Domain x/Domain y)
 7. Success after migration based on number of publications and citations.
 8. Hence using above info give proper suggestions for migration to author taken as input.

- In order to construct the GUI, Microsoft Visual Studio 2013 can be used. There Windows Form Application needs to be selected where we can design the dropdown boxes, search button, etc. Then a display would fetch the result from the appropriate program and display it. The programming language that has to be implemented for the above is C#.
- Similarly the features for Task 3, Task 4 and Task 5 also need to be collected, implemented using machine learning and the finally has to be embedded into the GUI.

OUTLOOK/CONCLUSION

In this project, we have studied the influence of personal interactions in a conference in the formation of the new citation and collaboration links. We have proposed a Graphical User Interface to perform some tasks based on the identified features for each task thereby provide information which is not easily available on the internet. We have also successfully created coauthor links for each domain and filtered it efficiently according to first publication of the authors. We have also performed a case study on the leading conferences in the "Networking and Distributed systems", "Artificial Intelligence", "Human-Computer Interaction" and "Hardware & Architecture" domain by collecting conference data from DBLP and using the citations file. We have identified the successful interactions for 3 different cases in the different domains and subsequently analyzed the properties of the induced citations. In our dataset, although the fraction of successful interactions which get translated into induced citation is quite low (9-12% max), nonetheless this conversion rate is rapidly increasing with time. This illustrates the fact that, as time progresses, authors become more aware about the benefit of the participation in a conference and interacting with the fellow researchers. We have also reported the influence of the successful interactions on the periodicity of the "induced" citation via toplayer. Thus we have proved the existence of such "induced" collaborations and analyzed their properties.

Thank You