"Whom-to-Interact": Does Conference Networking Boost Your Citation Count?

Soumajit Pramanik, Pranay Hasan Yerra, Bivas Mitra
Department of Computer Science and Engineering
Indian Institute of Technology Kharagpur, India
{soumajit.pramanik, ypranay.hasan, bivas}@cse.iitkgp.ernet.in

ABSTRACT

Recently, conference publications have gained a wide popularity, specially 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. 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 in future. We crawl the DBLP citation dataset and perform a case study on the leading conferences in the "Artificial Intelligence", "Hardware & Architecture", "Human-Computer Interaction" and "Networking & Distributed Systems" domains. Our empirical study is able to identify significant number of "successful" conference interactions which eventually results in "induced" citations. Interestingly, it is found that in most of the cases, it takes just 3 to 4 years to receive a citation from a participant interacted in a conference. It is also observed that the faster an interaction between two researchers can induce a citation between them, the longer this series of induced citations go on. Finally, we propose a machine learning based recommendation system 'Whom-to-Interact', for the researchers attending a conference, to suggest them 'with whom they should interact' for gaining incoming citations. The experimental results exhibit a decent performance of the system along with the impact of different regulating factors.

Categories and Subject Descriptors

D.2.8 [**SOFTWARE ENGINEERING**]: Metrics—*Performance measures*
; H.2.8 [**DATABASE MANAGEMENT**]: Database Applications—*Data mining, Scientific databases*
; H.3.4 [**INFORMATION STORAGE AND RETRIEVAL**]: Systems and Software—*Information networks*
; I.5.2 [**PATTERN RECOGNITION**]: Design Methodology—*Classifier design and evaluation, Feature evaluation and selection, Pattern analysis*
; I.5.5 [**PATTERN RECOGNITION**]: Implementation—*Interactive systems*

General Terms

Experimentation, Measurement, Performance

Keywords

Multiplex Networks, Citation Networks, Large Scale Network Evolution, Recommendation System

CODS '15 March 18 - 21, 2015, Bangalore, India.

http://dx.doi.org/10.1145/2732587.2732593.

1. INTRODUCTION

The citation network (between the academic articles) is a classic example of information network. It is now widely used to evaluate the impact of academic papers and the influence of scientists [6, 9]. Hence, the appearance of new citation links not only contributes to the topology of the network, but also has a tremendous impact on the scientific importance of a scientist which gets reflected by metrics such as h-index. In the last decade, several studies have been made in understanding and characterizing the citation network evolution [7, 1, 11]. One school of research has relied on the classical notion of preferential attachment and expects that the influential authors may attract the incoming citations preferentially [2]. On the other side, some work has been done on citation link formation instigating the topic migration of the researchers from one domain to another [8]. Again, in [10] citation network is also viewed from the perspective of information diffusion. In all cases, the existing approaches hardly provide any tool to the scientists to develop their own citations.

Like any other domain, scientific progress depends on the social communication and exchange of ideas. Initially the communication was mostly confined within the reviewing and referring to the existing literature published in various journal and archival publications. In this setup, most of the existing studies on citation network evolution are primarily focused on the dominant factors such as preferential affinity [5] towards the influential authors and topic migration of the researchers. However, those studies paid little attention to several dormant factors that play key roles in the evolution of the citation network. One such unexplored yet impactful dormant factor is *socialization of the researchers* in scientific conferences.

This is important to note that recent advancements and the popularity of the conferences, specially in the computer science domain, have opened up a new opportunity for the researchers to socialize with the fellow scientists. In these

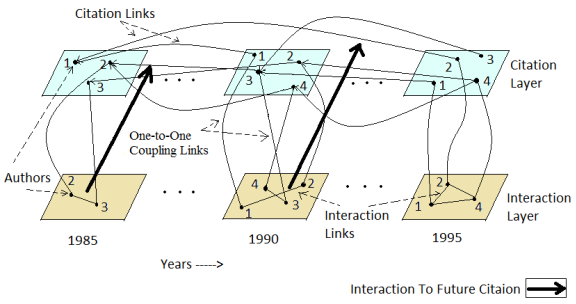


Figure 1: Multilayer Citation Network Representation

conferences, the participants get an opportunity to personally interact with the fellow researchers and exchange ideas. Most of these conferences are nicely structured (in tracks, sessions) facilitating the researchers of the similar domain to get familiar with each others work. In this context, this is very interesting to investigate that whether these social interactions (technical, non-technical) in the conferences eventually help to gain new citations; such interactions are termed as 'successful interaction'.

In a nutshell, a systematic study may eventually uncover the role of socialization on the formation of new citation links (micro level study) and evolution of the citation network (macro level study). This may have two direct implications; first of all, such study provides an external handle, in the form of the social interactions, to the participants/researchers, for improving their own influence in the community (by gaining new citation links). Secondly, further analysis may reveal the key factor behind the successful interaction (say, researchers' continent, affiliation etc) and the extent of the success (periodicity of citations, number of repetitions of citations etc).

Contribution: In this paper, we propose a multiplex network framework to uncover the influence of the participants' [1] interactions in a conference on the appearance of new citation links. In order to perform experiments, we crawl the citation and conference-session information during 1960-2008 from the DBLP repository and perform a comprehensive study on the four important domains of computer science - (a) "Artificial Intelligence" (b) "Hardware & Architecture" (c) "Human-Computer Interaction" and (d) "Networking & Distributed Systems". We show that this new representation enables us to identify the successful interactions at the conferences which eventually get induced into citations in the subsequent years. We introduce suitable metrics to realize the properties of these induced citations.

Our analysis identifies a significant amount of successful interactions and reveals various interesting properties of induced citations; for example fresh interactions have more influence in getting new citations than the ancient interactions. Moreover, the citation link induced through a recent interaction exhibits more recurrent behavior. The key factors behind the successful interaction and formation of induced citations, are explored. Finally, as an application of the framework, we propose a machine learning based model to predict the citation formation from the participants' interaction activities. We evaluate the performance of the model

[1]We use the terms 'Participant' and 'Authors' interchangeably in this paper.

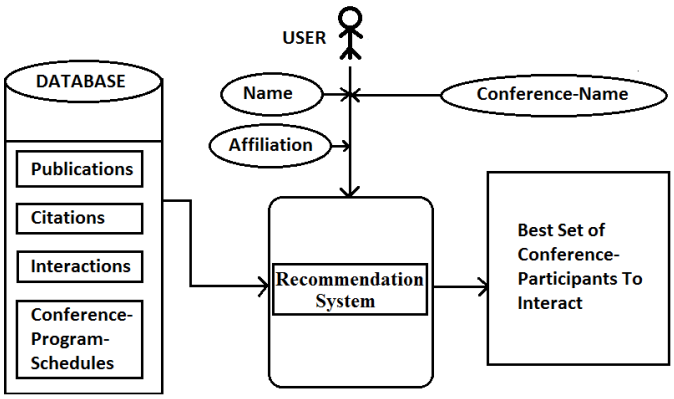


Figure 2: Block Diagram of the "Whom-To-Interact" Recommendation System

and show that it exhibits decent accuracy, precision and recall performance. The influence of the individual features on the performance of the model is also investigated. Finally, we outline the end to end recommendation system namely 'Whom-to-Interact' (see Fig 2), driven by the proposed machine learning based model. This system takes conference participant and conference information as input; as an outcome, it provides a ranked list of participants, with whom she may wish to interact in the conference, to increase her citation count.

The rest of the paper is organized as follows: Section 2 focuses on the dataset description. Section 3 proposes the methodology which revolves around the multiplex representation. In Section 4, we introduce the metrics and perform empirical study along with the insights. In Section 5, we develop and evaluate the machine learning model and outline the recommendation system 'Whom-to-Interact' . Finally in Section 6, we conclude the paper.

2. DATASET

In order to perform this study, basically we need to have two kinds of information (a) citation links among the authors (b) personal interactions between the conference participants.

We obtain the citation and collaboration information across the articles in computer science by crawling the DBLP dataset [3]. The dataset is current as of 2008. It is paper-centric: it describes 1 million different articles, back to year 1960, such that we know that a paper published in some year cites some other papers published in some possibly earlier year. This dataset contains the details of those 1 million research papers including their titles, author-names, publication-years, references and publication venues. We also tag the authors by their continents using "Microsoft Academic Search" utility. We transform this data into an author-centric dataset featuring which author cites whom in which year. The database contains a total of 6559415 citation links among the 501060 distinct authors of this 48-year collection. This citation data enables us to enumerate the evolution of the citation links starting from 1960 to 2008.

In order to gather the personal interaction information of the participants, we leverage on the publication venues information available in the above described dataset. In this paper, we focus on "Artificial Intelligence" domain, "Hard-

ware & Architecture" domain, "Human Computer Interaction" domain and "Networking and Distributed Systems" domain. We identify AAAI, ICDE, SIGIR and NIPS as the leading conferences in the "Artificial Intelligence" domain, CRYPTO, DAC and DATE as the leading conferences in the "Hardware & Architecture" domain, CVPR, ICIP and MM the leading conferences in the "Human-Computer Interaction" domain and INFOCOM, ICDCS, WWW and IPDPS as the leading conferences in the "Networking & Distributed Systems" domain. We crawl the program schedule of the aforesaid conferences (mainly from the conference portal, in some cases from DBLP sources) which essentially provides us the technical session information of the different conferences. The chosen time period for conferences is generally 1980-2007 and for those conferences which start after 1980, the time-period is from its starting year to 2007. The choice of the time-periods are mainly driven by the availability of data. A quick glimpse to the collected data reveals that each conference has on average 45-50 sessions (including parallel tracks) where in average 3-5 papers are presented in each technical session. This session information eventually provides us the interaction statistics between the participants in the conference.

3. METHODOLOGY

Our methodology relies on a temporal multiplex network representation, capturing the citation information between the authors and the participant interactions in a conference [Fig. 1]. For each year, we construct a multiplex network with two layers; The top-layer contains all the authors (as nodes) who published a paper in that year (say t_A). A directed citation link connects author A in year t_A with author B (of year $t_B \geq t_A$) if in her paper, author B cites A. The bottom layer of the multiplex contains the (author) participants in a conference as nodes and the interaction between the authors as links. As mentioned earlier, in this study, we restrict ourselves within the chosen conferences in the field of "Artificial Intelligence", "Hardware & Architecture", "Human-Computer Interaction" and "Networking & Distributed Systems". Since the interaction information between two authors is not readily available, we take the help of following two realistic assumptions (a) The authors, whose talks are scheduled in the same technical session of a conference, have high chances of interaction. (b) In general, the first or the last author (or sometimes both) of a paper attends the conference. We have verified this assumption using the data from the portals of two conferences- CIMTA (Computational Intelligence: Modeling, Techniques and Applications) 2013 and NCC (National Conference on Communications) 2014 where the information about the registered participants are available along with the program schedule. In the next section, we portray our results with different probabilities of first/last/both author(s) attending the conference and then we point out that the claims made in the results are independent of the specific choice of these probability values.

Since the technical session information of most of the conferences are readily available, we can easily construct the lower layer of the multiplex network by forming the interaction links based on the aforesaid assumptions. Once the top and bottom layer has been constructed, we couple the two layers using the one-to-one authorwise links (see Fig. 1). Thus, a suite of timestamped multiplex network is constructed for each year from 1980 to 2008 to understand the network evolution. Once the network suite is constructed, we are ready to investigate the impact of interactions at the bottom layer on the appearance of citation links on the top layer.

3.1 Successful Interaction

Informally, if an interaction between two participants results in a new citation, we designate that as a *successful interaction*. More precisely, interaction between two participants x and y at time t (at the bottom layer of the multiplex) is considered '*successful*' for participant x, if (a) this interaction leads to the creation of an incoming citation link from y to x at time $t+1$ onwards (at the top layer of the multiplex) and (b) there does not exist any citation or collaboration (co-authorship) between node pair x and y before time t. Participant x is termed as '*successful*' with respect to this interaction. The citation links appeared as a result of interactions where at least one of the participating authors is '*successful*' are defined as '*induced*' citations.

Impact of Research groups: Interestingly, working in research groups are gaining wide popularity these days. These research groups manifold the impact of the conference interactions. For instance, participants, working in research groups may percolate the interaction information to her own group members. Here essentially through just one interaction between two participants, two research group members may get familiar with other's work. This is termed as 'group interaction'. Hence, one *successful* interaction between two participants may result in *induced* citations between several 'non-interacting' (group) members. We term that as the *successful* 'group interaction'. Figure 4 illustrates the process. Note that the group ($Group_1$ in Fig 4) is termed as '*successful*' if at least one of the authors (say D) of that group is able to receive a citation from at least one of the members (say Q) of any other group (say $Group_2$).

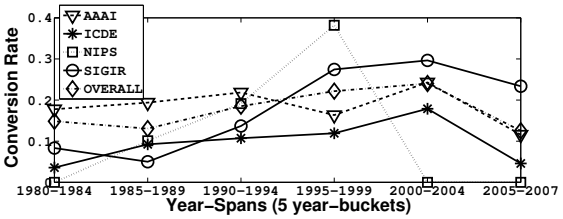
Accurately identifying the research groups over a time period is a separate research problem, which is outside the scope of this paper. Nevertheless, we concentrate on article-based groups which assumes that all the co-authors of an article are part of the same 'virtual' research-group at the time of publication of the paper. This enables us to identify the *successful* 'group interactions'.

4. EMPIRICAL STUDY

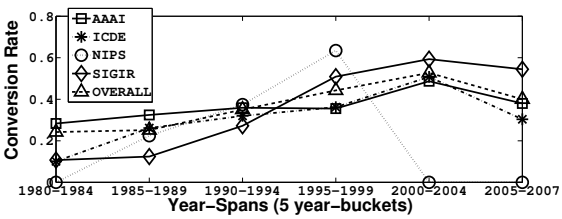
We start this section by defining a set of metrics used to characterize the important properties of the induced citation links. Next, we identify the influence of personal interactions on citation formation and subsequently explore their properties. In our work, we check the impact of an interaction on the persons participating in it as well as their groups. So, we define our evaluation metrics for both the cases separately in the following subsection. First we define the metrics we use to analyze the effect of interaction on individual's career and then we extend the same metrics for analyzing the effect on their groups' prospect.

4.1 Evaluation metrics

In this section, we introduce metrics to characterize the properties of the *successful* interactions and *induced* citations. Note that while developing each metric, we distinguish the individual interaction (I) and group interaction (G).



(a) Author-wise Conversion Rates (I_{CR}) for Overall and each conference in AI domain.



(b) Group-wise Conversion Rates (G_{CR}) for Overall and each conference in AI domain.

Figure 3: Author-wise (I_{CR}) & Group-wise (G_{CR}) Conversion Rates for AI Domain

1. Conversion Rate I_{CR} (G_{CR}): The conversion rate measures the propensity of an author being '*successful*' in converting a new interaction in a conference to a new induced citation link. Formally, we define conversion rate for a conference as the ratio between the number of '*successful*' participants (A_S) and total number of participants of that conference (A_T) (except those authors who have cited or collaborated with authors from their own sessions in past).

$$I_{CR} = \frac{A_S}{A_T} \tag{1}$$

From this, the definition of the domain-wise overall conversion rate can be simply extended by taking the ratio of the sum of all the '*successful*' authors attending any conference in a domain and total number of participants of all the conferences in that domain. Conversion rate realizes the influence of personal interactions on the appearance of citation links and plays a key role behind the claims made in this paper. Extending this concept for group interactions, we define group based conversion ratio (G_{CR}) as the fraction of '*successful*' groups out of the total groups in that conference (or domain) during that time-period.

2. Induced Citation Link Repetition Count I_{LR} (G_{LR}): Once an '*induced*' citation link is formed as a result of the '*successful*' (group) interaction, '*Induced Citation Link Repetition Count*' measures the recurrent appearance of the link in the recorded time period.

3. Lifespan of Induced Citation I_{LS} (G_{LS}): The Lifespan of the '*induced*' citation is measured as the difference between the first and last appearing year of the '*induced*' citation link.

4. Influence Gap of Successful Interaction I_{IG} (G_{IG}): The influence gap of a '*successful*' interaction is measured as the latency between the '*successful*' (group) interaction and the formation of the first '*induced*' citation.

4.2 Influence of Interactions on Citation

In this section, we identify the 'successful' authors observed in the "Artificial Intelligence", "Hardware & Architecture", "Human-Computer Interaction" and "Networking & Distributed Systems" domains and then characterize the properties observed in the corresponding "induced" citations.

4.2.1 Conversion rate

We start with the computation of conversion rate for different domains. In this line (as mentioned in the previous section), we take different possibilities regarding the participation of the first/last/both author(s) in the conferences. To be precise, we use 3 sets of probabilities to calculate the conversion rates - case (a) $[0.7, 0.3]$, case (b) $[0.8, 0.2]$ and case (c) $[0.9, 0.1]$ where the first element depicts the (probability of) presence of either the first or the last author and the second element shows the (probability of) presence of both of them. In general, it is little unlikely to have both the first and the last author to present the paper, hence we keep that probability low. The sets of probabilities observed in CIMTA'13 and NCC'14 are respectively [0.88,0.12] and [0.68,0.32] which fit well within the set of probabilities chosen by us. This author participation statistics enables us to construct the the interaction (bottom) layer of the multiplex network.

Observation: In the chosen conferences of the "Artificial Intelligence" domain, we observe a conversion rate (I_{CR}) 16.8% (2223 out of 13234 authors) for case (a). The conversion rate becomes 16% (2088 out of 13085 authors) for case (b) and 15.1% (1949 out of 12921 authors) for case (c). So, it is quite evident that the interaction probabilities do not affect the conversion rates much. So, in the rest of this paper we report the results of the experiments considering case (b) [0.8,0.2]. Performing similar experiments for "Hardware & Architecture", "Human-Computer Interaction" and "Networking & Distributed Systems" domains give us conversion rate of 15.2% (1262 out of 8282 authors), 8.9% (1294 out of 14549 authors) and 5.7% (713 out of 12550 authors) conversion rates respectively.

We also calculate the conversion rate (G_{CR}) for group-wise interactions for the same domains. We find the conversion rate 39% (3790 out of 9709) for "Artificial Intelligence" domain, 42.5% (2195 out of 5166) for "Hardware & Architecture" domain, 24% (2620 out of 10926) for "Human-Computer Interaction" domain and 22% (1876 out of 8458) for "Networking & Distributed Systems" domain. In Table 1, we report the group-wise as well as the author-wise conversion rates of each domain. It shows that the group-wise conversion rates are around 2-3 times more than the author-wise conversion rates. This indicates that if two participants interact in a conference, even if they do not cite each other, there is a high possibility that the group-members may gain citations.

Claim and the evidence: Albeit the absolute magnitude of the 'successful' authors is not very high, nevertheless, this straightaway indicates that this first interaction becomes the key factor behind the formation of new citation link between the conference participants. In order to establish our claim, we perform the following experiments to show that the appearance of the new citation links on the top layer indeed results from the participants' interaction; this is neither a statistical fluctuation nor from a sporadic

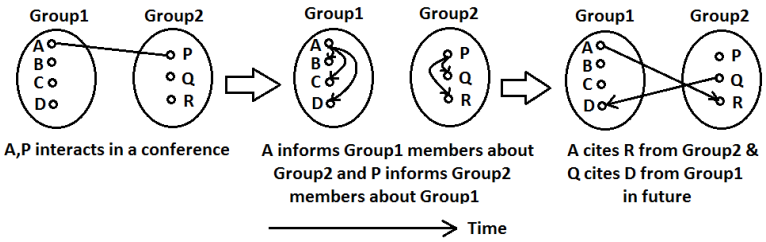


Figure 4: Example showing how an interaction between two persons can create citations between their other group-members

Domain	Group-wise Conversion Rate	Author-wise Conversion Rate
Artificial Intelligence	39%	16%
Hardware & Architecture	42.5%	15.2%
Human Computer Interaction	24%	8.9%
Networking & Distributed Computing	22%	5.7%

Table 1: Group-wise & Author-wise Conversion rate for different domains

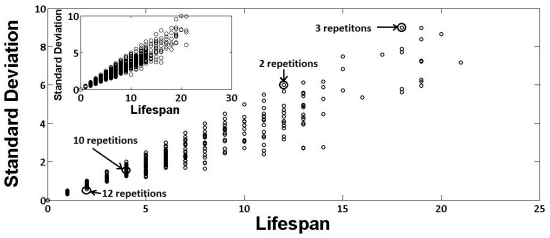


Figure 5: Plot of the Standard deviations of the repetition years vs Lifespans (I_{LS}) in AI domain for Author-wise interactions (Inset shows the similar statistics in AI domain for Group-wise interactions)

effect. We remove the 'induced' citations links from the top layer of the multiplex network and replace them using the following three strategies

1. Random Replacement: In this case, we replace 'induced' citations by citation links between any two randomly chosen authors from the author set.

2. Replacement by random authors of same year: We replace each 'induced' citation by a citation link between two random authors appeared in the same year as the respective 'successful' authors.

3. Replacement by successful authors: We replace each 'induced' citation by a citation link between two randomly chosen 'successful' authors.

After replacing the 'induced' citations, we recalculate the conversion rates in each aforementioned case. Interestingly, we do not find a single 'successful' author for any of the replacement strategies. This clearly indicates that the existence of the 'induced' citations is not a random or sporadic event; they are definitely the result of those conference interactions where at least one participating author is 'successful'.

4.2.2 *Distinguishing Properties of Successful Authors*

Next, we aim to identify the underlying properties of 'successful' authors which differentiates them from the others. We find the following two discriminating properties.

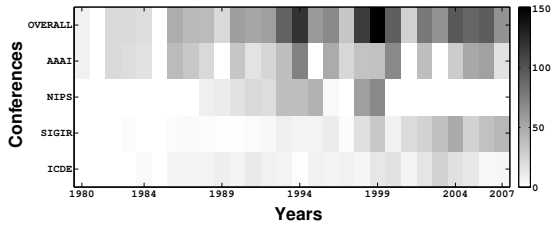
1. **Interaction Count:** Interaction count of an author can be defined as the number of times an author is involved in interactions in conferences. It is observed that the average interaction count of the 'successful' authors is 10.7 whereas this average is 4.7 for the 'unsuccessful' authors. This clearly points to the fact that the 'successful' authors more aggressively interact with the conference participant than others.

2. **In-Citation Count:** This is observed that the 'successful' authors have on average high in-citation count (488) against the other authors (136). This indicates that successful authors mostly works as an *authority* in the research community. Additionally, conference interaction helps them to gain new incoming citations, compared to the others.

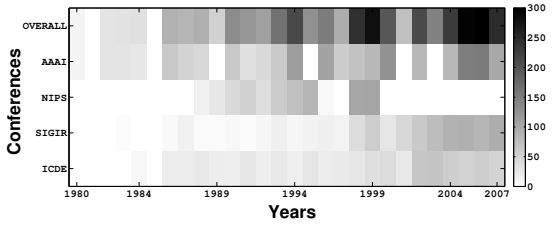
In order to avoid repetitions, we primarily display the results of the "Artificial Intelligence" domain in the subsequent sections.

4.2.3 *Evolution of conversion rate with time*

In this section, we illustrate how does the conversion rates I_{CR} and G_{CR} evolves over the years for different conferences. In Fig. 3(a) and Fig. 3(b), we plot the author-wise and group-wise conversion rates (I_{CR} and G_{CR}) for the individual conferences as well as the evolution. From the plots, it is evident that in "Artificial Intelligence" domain the overall and conference-wise conversion rates mostly increase from 1980 to 2005 (noticeably, G_{CR} increases at a higher rate than I_{CR}), which indicates that with time, people are gradually becoming aware of the utility of interactions during the conferences. Only for 'NIPS' conference, we see a sharp fall after 1999 which is due to the scarcity of information after 1999.
Interestingly, we notice that at the last year-bucket (i.e. between 2005 to 2007), all the conversion-rates drop simultaneously. Apparently it might appear that over time, influence of personal interaction on citation formation gradually diminishes; however a more closer look uncovers the true dynamics behind this fall. We find that on average it takes around 3 to 4 years (3.4 to be exact) to get the first "induced" citation link from a successful interaction. As a result, we observe a fall after the year 2005 in both I_{CR} and G_{CR} since



(a) Author-wise AI domain



(b) Group-wise AI domain

Figure 6: Heatmaps showing conference-wise contributions to successful interactions in AI Domain

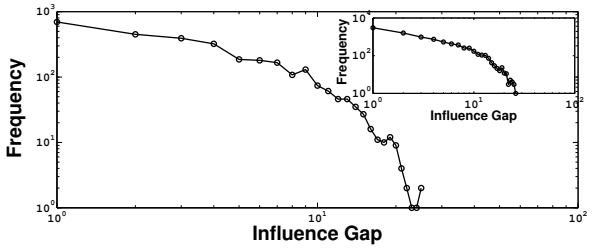


Figure 7: Frequency Distribution of Author-wise Influence Gaps (I_{IG}) in AI domain (Inset shows the similar statistics for Group-wise Influence Gaps (G_{IG}))

the interactions during 2005-2007 do not get enough time to generate "induced" citations. We term this as *boundary effect*.

In the above discussions, we primarily focus on the conversion rate which is essentially a relative quantity. In Fig. 6(a) and Fig. 6(b), we plot the heatmaps to show the conference-wise contribution of the absolute number of "successful" authors per year. Concurring with the intuition, in "Artificial Intelligence" domain, we observe that the absolute values during the initial years are quite low. However, from 1990 onwards the overall count gradually increases, thanks to the popularity of 'AAAI' conference. Though, throughout we see a dominance of 'AAAI' conference over others in contributing "successful" authors, 'NIPS' also contributes significantly over the time-period 1988-1999 during which the session information of 'NIPS' is available to us.

4.2.4 Periodicity of Induced citations

Next, we examine the (re)appearances of the individual "induced" citations over the year. Essentially here we investigate the nature of periodicity of the appearance of those induced citation links in their entire lifespan. We plot the periodicity (measured by standard deviation) profile of each inducted citation with respect to the lifespan in Fig. 5 (for "Artificial Intelligence" domain). As the standard deviation linearly increases with the lifetime, it implies that the repetitions of the induced citations are not very periodic and their skewness increases with the lifetime of the repetitions. We mark few interesting anecdotes in Fig. 5. Here we point to few induced citations having the high repetition count (I_{LR}) but low standard deviation, showing recurrent behavior; on the other hand, few induced citations have low repetition count but high standard deviation, exhibiting sporadic appearance.

4.2.5 Influence of Successful Interactions

So far, we have concentrated on the different characteristics of the induced citation links. Now we turn our attention to the "successful" author-wise (or group-wise) interactions and assess their influence on the formation of new citation links. We measure the influence gap of interaction (I_{IG} and G_{IG}) as the latency between the year of "successful" interaction and the appearance of first induced citation. Detail analysis follows.

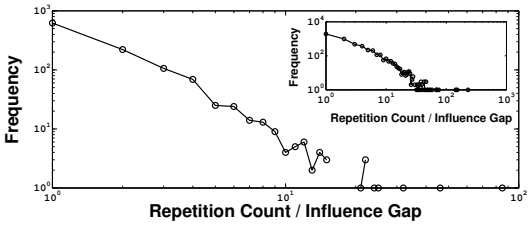
1. Influence Gap Distribution: We start with the distribution of I_{IG} and G_{IG}. From the frequency distribution in Fig. 7, it becomes visible that most of the "induced" citations have very small "influence gap". This points to the fact that in most of the cases, personal interaction quickly turns into a new citation. This is also an evidence that conference interactions may indeed induce future citations; a large fraction (70%) of "successful" authors receive the first "induced" citation just within the 5 years of the interaction, which cannot be just a mere coincidence.

2. Influence Gap and Sustainability: Next we examine how the "influence gap" i.e. I_{IG} and G_{IG} of the "successful" interactions affect the sustainability of the "induced" citations, measured by their repetition counts I_{LR} and G_{LR}. First we plot the frequency distribution of the repetition count influence gap ratio (I_{LR}/I_{IG} and G_{LR}/G_{IG}) [see Fig. 8(a)] for "Artificial Intelligence" domain. This figure shows that most of the "induced" citations have low value of this ratio. To observe the influence gap and the corresponding repetition counts of different induced citations in more detail, we place them in a repetition count vs influence gap plane [Fig. 8(b)].

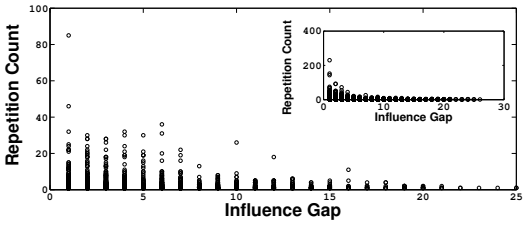
Observation: In Fig. 8(b) we show that a large population of citations accumulate at the left side of the figure. This conveys two important messages - firstly, most of the "induced" citation have highly influential interactions with short influence gaps which result in achieving a high repetition count and secondly, if for an "induced" citation the "influence gap" is very long, it is very rare that the citation link have a high repetition count.

4.2.6 Impact of Author Continent

Finally we investigate the influence of the continents of interacting authors on the formation of the citation links between them. In our dataset, we have authors from 5 different continents - Asia, Africa, Europe, North America and South America. In Fig. 9(a), we show the percentage of 'successful' authors from each continent for each domain. We observe that, consistently, authors from 'North America' are the most 'successful' authors in any domain. If we rank the continents in descending order based on the population of

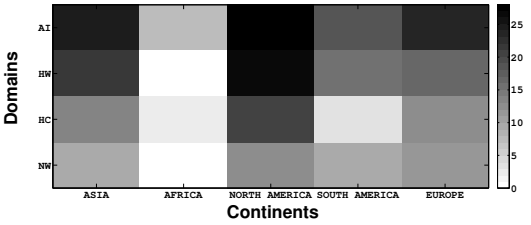


(a) Frequency distribution of the ratio of Repetition Count (I_{LR}) and Influence gap (I_{IG}) in AI domain (Inset shows the similar statistics for Group-wise Repetition Count (G_{LR}) and Influence gap (G_{IG})).

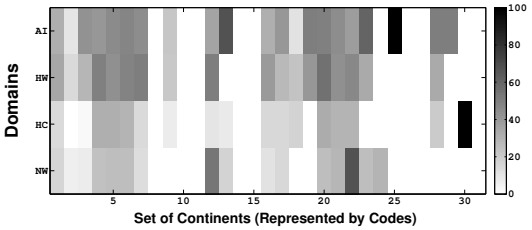


(b) Author-wise Repetition counts (I_{LR}) vs Influence gaps (I_{IG}) in AI domain (Inset shows the similar statistics for Group-wise Repetition counts (G_{LR}) vs Influence gaps (G_{IG})).

Figure 8: Impact of Influence Gap (I_{IG} & G_{IG}) on Repetition Count (I_{LR} & G_{LR}) in AI Domain



(a) Heatmap of conversion rates (I_{CR}) (scaled by 100 for plotting) depending on the continents of the participating authors in AI domain.



(b) Heatmap of conversion rates (G_{CR}) (scaled by 100 for plotting) depending on the sets of continents of the authors belonging to the participating groups.

Figure 9: Heatmap showing conversion rates for individual authors and author-groups for different continents

'successful' authors, we find the following sequence- 1) North America 2) Asia 3) Europe 4) South America and 5) Africa.

Showing the similar results for group-wise interactions is relatively difficult because each group contains authors from multiple continents. In order to represent a subset of the 5 continents as a code between 1 to 31, we use the following coding technique.

$Code = 16 \times P_{Asia} + 8 \times P_{Africa} + 4 \times P_{NorthAmerica} + 2 \times P_{SouthAmerica} + 1 \times P_{Europe}$

where P_C=1 if there is at least 1 author from continent C ($C \in$ {Asia, Africa, Europe, North America, South America}) in the group and P_C=0 otherwise. For example, if the code of a group is 21 (=16+4+1), it implies that the authors are from Asia, North America and Europe. Now, using the above code, we plot the percentage of "successful" authors from each set of continents for each domain [Fig. 9(b)]. The best percentage of "successful" authors is observed for the group with code 22 (i.e. Asia, North America & South America). Groups with codes 4 (i.e. only North America), 5 (i.e. North America & Europe), 6 (i.e. North America & South America), 20 (i.e. Asia & North America) and 21 (i.e. Asia, North America & Europe) have also shown good success-rates. If we consider the groups where all authors are from the same continent (i.e. code 1,2,4,8 & 16), the group with code 4 (i.e. only North America) has done the best. Groups with codes 1 (i.e. only Europe) and 16 (i.e. only Asia) have also performed reasonably well.

5. RECOMMENDATION SYSTEM

Finally, in this section we aim to develop a recommendation system namely 'Whom-to-Interact' (see Figure 2). The proposed system can recommend a scientist the possible way, in which she can gain new incoming citations. For example, participating in a conference and interacting with the fellow scientists gives good opportunity for getting new citations. The proposed 'Whom-to-Interact' service may provide a suitable guideline to the conference participant, with whom she may try to interact during the conference, in order to increase her citation count. More specifically, given a conference C and an user U, this system provides a suggested list of participants for interaction. These participants are ranked based on the propensity that they cite the user U in near future.

5.1 Proposed Model

The core of the recommendation system is a supervised machine learning based model. This model infers the citation formation from the participants' interaction activities. In the following, we build the inference feature table based on the users past citation activities. The empirical study in section 4 uncovers the important factors regulating the citation formation as a result of interactions. Hence, in the feature table (Table 2), we focus on the three specific categories of features (1) citation record (2) interaction record (3) co-authorship record.

1. Citation record: Collectively, citation count and publication count of author pairs (in different forms) are important factors behind formation of citations. Specifically, we compute (i) total citation-counts of the author pairs (total impact), (ii) difference of citation-counts of author pairs (differential impact), (iii) difference in publication-counts (difference in Experience) (iv) count of mutual citations (Sim-

Feature Categories	Feature-Index	Features
Citation Record	(1)	Sum of citation-counts of author-pairs
	(2)	Difference of citation-counts of author-pairs
	(3)	Difference of publication-counts of author-pairs
	(4)	Sum of mutual citations of author-pairs
	(5)	Minimum of mutual citations of author-pairs
Interaction Record	(6)	Sum of conversion-rates of author-pairs
	(7)	Multiplication of conversion-rates of author-pairs
	(8)	Sum of successful-interaction-counts of author-pairs
	(9)	Multiplication of successful-interaction-counts of author-pairs
Co-Authorship Record	(10)	Number of common co-authors
	(11)	Number of times participating authors co-authored
	(12)	Continents of the participating authors (each continent-pair represented by a unique code)

Table 2: Features used for Recommendation System

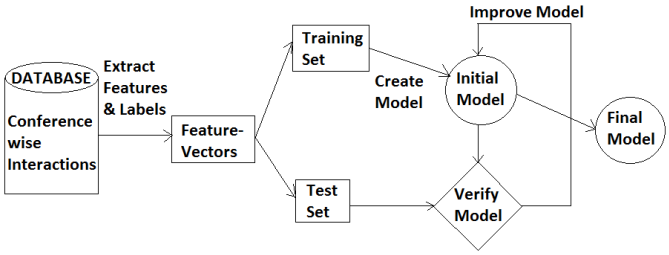


Figure 10: Supervised Learning Model for Recommendation System

ilarity of research-areas).

2. Interaction record: We use different forms of interaction information and conversion statistics, which are essential for the citation prediction. Precisely, the addition and multiplication of the conversion rate and the number of successful interactions between the participants in a conference carry a strong signal for the successful future interaction. Moreover, the time (year) gap (say n) between the participant interaction and formation of first citation is also an important feature to predict the year of first citation after the interaction.

3. Co-Authorship record: Co-authorship record points to the usage of a co-authorship information to predict the induced citation. This includes identifying the common set of co-authors between the pair of participants, co-author count etc. The information about the continents of author-pairs is also an important feature in this category.

5.2 Performance Evaluation

In this section, we implement and evaluate the performance of the proposed machine learning model. For each conference, we randomly choose a sufficiently large set of evenly balanced successful and unsuccessful interactions. Since our dataset has information about the interactions and citations between 1980-2008, we use the features of the interactions up to 1998 and predict the citations between 1999-2008. As we already have the citation data between 1999 to 2008, we can easily validate the correctness of the prediction.

We use 75% of the total interactions in each conference occurred during 1980-1998 as "training set" to train our "Support Vector Machine" classifier and the remaining 25% of these interactions as "test set" to test its performance. First, we systematically extract the 12 features from each such chosen interaction and scale them accordingly(see Figure 10). We tag each feature vector with a class label 1 or 0 depending on whether the corresponding interaction is successful (induce an incoming citation within next n years ($1 \leq n \leq 10$)). We use the standard "libsvm-3.18" package [4] to implement the model.

5.2.1 *Evaluation metrics*

We define the metrics to evaluate the performance of the model. The metrics are based on the following four sets.

True Positive (TP) set : The "successful" interactions in the test-set which are also predicted to be "successful" by the model.
True Negative (TN) set : The "unsuccessful" interactions in the test-set which are also predicted to be "unsuccessful" by the model.
False Positive (FP) set : The interactions in the test-set which are predicted to be "successful" by the model but actually "unsuccessful"
False Negative (FN) set : The interactions of the test-set which are predicted to be "unsuccessful" by the model but actually "successful".
Next, we define the evaluation metrics.

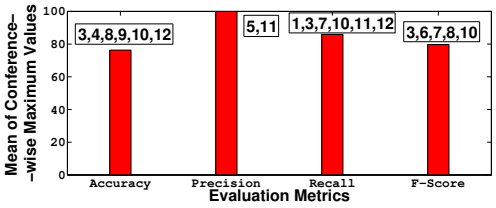
(1) $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$
(2) $Precision = \frac{TP}{TP+FP}$
(3) $Recall = \frac{TP}{TP+FN}$
(4) $F_1 Score = 2.\frac{Precision \times Recall}{Precision+Recall}$

5.2.2 *Experimental Results*

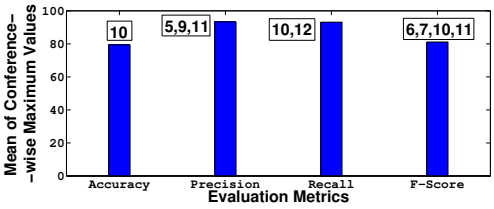
This is indeed comforting for us to observe a decent performance of the recommendation model (see Fig. 11). For each conference , we observe on average 80% accuracy, 98% precision, 91% Recall and 81% $F_1 - Score$. Next we investigate the role of the different features on the performance of the system in the various domains.

Feature Analysis

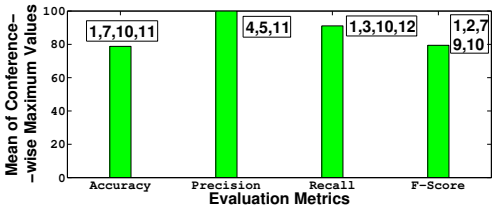
First, we identify the most influential features of each individual domain. In Fig. 11, we show the performance of the model in each domain along with the most important features (with respect to individual metrics). Summarizing, in Table. 3 we tag each domain with the set of features which generate the best results for all the metrics in that



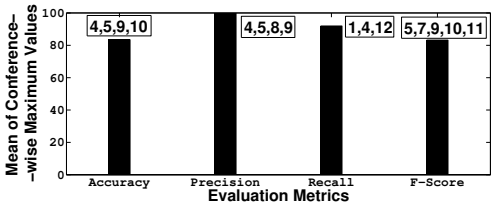
(a) Evaluation Metrics' Values in AI domain



(b) Evaluation Metrics' Values in HW domain



(c) Evaluation Metrics' Values in HC domain



(d) Evaluation Metrics' Values in NW domain

Figure 11: Evaluation Metrics' Values for Different Domains

domain. We observe that the feature 10 (the number of common co-authors of the interacting participants) to be the best performing feature for all the domains. Apart from feature 10, features 7 (Interaction Record Category) and 11 (Co-authorship Record Category) also show quite good performance for all the domains. However, features from Citation record category work well only for few specific domains.

Similarly, we identify the features that perform well for individual metrics. In Table. 4, we tag each metric with the set of most important features. Here the features 4, 9, 10 and 11 are found to perform well for all the metrics.

5.2.3 *Generality and Scalability of the Model*

The performance reported till now are calculated using conference specific features, i.e. we train and test the model for a specific conference. But in reality, conference specific model may not be always feasible as it is expensive and time critical to train the model for each conference. For example, in case of new conferences, the size of the training and test set are insufficient to train the model. Moreover, it needs to repeat the computation for each conference which makes the system less efficient. Here we would like to explore the possibility, if we can use the domain specific model (building feature table from the selected conferences in a domain) or a single generalized model (building feature table from all the conferences) for prediction.

Intuitively, it appears that domain specific model or single general model should perform quite poorly in comparison with conference-wise model. To verify that, we use the same features to create domain-wise models & a unified single model and evaluate their performance (see Fig. 12). Interestingly, none of the metrics exhibit major changes. The accuracy proves to be even better for the single model than conference-wise models. One possible reason, the training and test sets of the single general model are substantially large and comprehensive (as we put the information of all conferences), compared to conference-wise models and this helps the system to learn better. Moreover, the successful

Domain	Feature-Set
Artificial Intelligence	3,7,8,10,11,12
Hardware & Architecture	10,11
Human Computer Interaction	1,7,10,11
Networking & Distributed Computing	4,5,9,10

Table 3: Domain-wise Best Features

Metrics	Feature-Set
Accuracy	4,9,10
Precision	4,5,9,11
Recall	1,3,10,12
F-Score	6,7,9,10,11

Table 4: Metric-wise Best Features

interaction is a very generic property and does not change frequently with each conference. This also proves the robustness of the features we have chosen for our model. Summarizing, in order to boost the performance, it is possible to switch to domain-wise or single general model without compromising the performance.

5.3 Outlook

Finally, we illustrate how our model can be used to develop the "Whom-to-Interact" recommendation system (see Figure 13). The system has two components, namely the front end *aka* user interface as well as the back end. The back end of the system implements the proposed machine learning model. In the front end, the user (say U) will be asked for her name, affiliation and the conference she wish to attend. The back end of the system crawls the "Program Schedule" of that conference to get the participant list and figures out the list of participants whose talks have been scheduled in the same or non overlapping sessions. Next, for each such participant X, the system creates a pair (U, X) and predicts whether the interaction will lead to a citation for U in the next n years (n can vary from 1 to 10); it ranks the participants based on the probabilities. Finally, the user

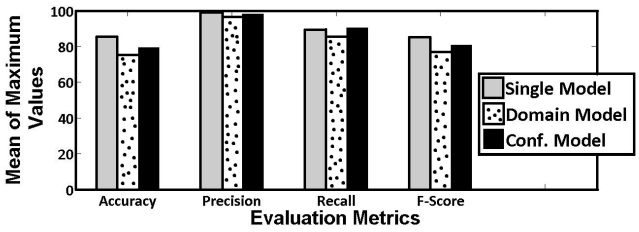


Figure 12: Evaluation Metrics' Values for Different Types of Models

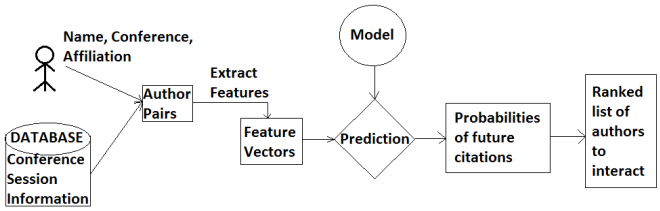


Figure 13: Prediction of future citation using Recommendation System

U can choose the authors from the ranked list she wishes to interact in the conference to gain more incoming citations.

6. CONCLUSION

In this paper, we have made an important contribution in explaining the role of social communications on the career and academic importance of a researcher. We have studied the influence of personal interactions in a conference on the formation of the new citation links. We have proposed a multiplex network framework to represent the DBLP citation dataset and performed a case study on the leading conferences in the "Artificial Intelligence", "Hardware & Architecture", "Human-Computer Interaction" and "Networking & Distributed Systems" domains. We have identified a significant fraction of successful interactions in the different domains (specially in case of group interactions, $G_{CR} \approx 43\%$) and subsequently analyzed the properties of the induced citations. Fig. 6(a) clearly reveals that this conversion rate is rapidly increasing with time. This illustrates the fact that, as time progresses, authors become more aware of the benefit of the conference-networking with the fellow researchers. Moreover, the quick formation of the first citation link, as a result of successful interaction, leaves a more persistent and long standing effect on the future successive citations. Our analysis also revealed that interaction between the 'North American' participants proves more beneficial for attracting new citations, where as group-wise interaction between the participants from North America, Asia and South America attracts more citations. Based on several identified features, we have proposed a machine learning model to predict the future citations from the past interaction. The evaluation experiments confirm that the model exhibits a decent performance (with high accuracy, recall, precision value). Model analysis has shown that some of the important features such as co-authorship record (of the interacting participants) have strong correlation on the predicted citation. Finally as an application of the model, we have outlined a recommendation system 'Whom-to-Interact' which may help the participant to decide, with whom she should interact in the conference to gain more citations.

7. ACKNOWLEDGMENTS

This work has been partially supported by the SAP Labs India Doctoral Fellowship program and DST - CNRS funded Indo - French collaborative project titled "Evolving Communities and Information Spreading".

8. REFERENCES

[1] F. Amblard, A. Casteigts, P. Flocchini, W. Quattrociocchi, and N. Santoro. On the temporal analysis of scientific network evolution. In *CASoN*, pages 169–174, 2011.

[2] A. L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert, and T. Vicsek. Evolution of the social network of scientific collaborations. *Physica A: Statistical Mechanics and its Applications*, 311(3-4):590 – 614, 2002.

[3] T. Chakraborty, S. Sikdar, V. Tammana, N. Ganguly, and A. Mukherjee. Computer science fields as ground-truth communities: Their impact, rise and fall. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13, pages 426–433, New York, NY, USA, 2013. ACM.

[4] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software Available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

[5] D. J. de Solla Price. Networks of scientific papers. *Science*, 149(3683):510–515, 1965.

[6] D. Dieks and H. Chang. Differences in impact of scientific publications: Some indices derived from a citation analysis. *Social Studies of Science*, 6(2):pp. 247–267, 1976.

[7] Y.-H. Eom and S. Fortunato. Characterizing and modeling citation dynamics. *PLoS ONE*, 6(9):e24926, 09 2011.

[8] Q. He, B. Chen, J. Pei, B. Qiu, P. Mitra, and C. L. Giles. Detecting topic evolution in scientific literature: how can citations help? In *CIKM*, pages 957–966, 2009.

[9] E. A. Leicht, G. Clarkson, K. Shedden, and Newman. Large-scale structure of time evolving citation networks. *The European Physical Journal B - Condensed Matter and Complex Systems*, 59(1):75–83, 2007.

[10] X. Shi, B. L. Tseng, and L. A. Adamic. Information diffusion in computer science citation networks. In *ICWSM*, 2009.

[11] L. Šubelj and M. Bajec. Model of complex networks based on citation dynamics. In *Proceedings of the 22Nd International Conference on World Wide Web Companion*, WWW '13 Companion, pages 527–530, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.