Analysing Evolution of Sleeping Beauty In Heterogeneous Citation Network

A PROJECT REPORT
SUBMITTED IN PARTIAL FULFIMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
Master of Engineering
IN
Computer Science and Engineering

BY

Siddharth Shakya

Computer Science and Automation
Indian Institute of Science
Bangalore – 560 012 (INDIA)

June, 2016

Declaration of Originality

I, **Siddharth Shakya**, with SR No. **04-04-00-10-41-14-1-11174** hereby declare that the material presented in the thesis titled

Analysing Evolution of Sleeping Beauty
In Heterogeneous Citation Network

represents original work carried out by me in the **Deparment of Computer Science and Automation** at **Indian Institute of Science** during the years **2014-2016**.
With my signature, I certify that:

- I have not manipulated any of the data or results.
- I have not committed any plagiarism of intellectual property. I have clearly indicated and referenced the contributions of others.
- I have explicitly acknowledged all collaborative research and discusions.
- I have understood that any false claim will result in severe disciplinary action.
- I have understood that the work may be screened for any form of academic misconduct.

Date: Student Signature

In my capacity as supervisor of the above-mentioned work, I certify that the above statements are true to the best of my knowledge, and I have carried out due diligence to ensure the originality of the report.

Advisor Name: Advisor Signature

DEDICATED TO

The Student Community

who can use and reuse this template to glory

Acknowledgements

I express my gratitude and sincere thanks to Prof. M Narsimha Murty for providing the opportunity to work, on an interesting but lesser studied problem "Evolution Of Sleeping Beauties In Science Literature", under his guidance that encouraged critical thinking which was necessary for a project like this and being supportive throughout the project. I also want to thank all the lab mates for being generous enough in giving their time for fruitful discussions and honest feedback, without which I might have missed out on some subtle points that were very helpful in the project. I thank Govind Sharma for introducing me to this problem and providing useful insights much needed for this project, Sharad Nandanwar for helping me with the mathematical aspect of the problem and Anubhav Gupta for helping me with python and debugging my codes.

I would also like to thank IISc for creating and maintaining a wonderful environment for carrying out research.
Last but not the least I would like to thank my family members for providing their support when I needed it the most.

Abstract

Heterogeneous citation network consists of different types of members having different characteristics, involved in interacting and influencing each other in different ways. With time, the nature of interaction amongst the members in the network changes and these changes act as the indicators for the direction in which a community is heading to. Objective of our project is to analyse the evolution of sleeping beauty literature in citation network to find the factors that effect their evolution and quantify their effect. We identified sleeping beauties in the Pubmed and computer science subset of Microsoft Academic Network Dataset using a recently published technique. We observed that the delayed increase in their citation count is not sudden but gradual, hence we partitioned their delayed citation accumulation phase into waking period and peaking period. Our analysis shows that citing authors' rank has high correlation with citations accumulated in its peaking period but not in waking period. Rank Pattern Analysis showed an intense increase in citation count around the time an influential author cites it. To explain the observations of rank pattern analysis we introduced a novel notion of followership in citation network. We proposed a method to calculate following coefficient for quantifying the followership strength of an author with respect to another author. Follower Pattern Analysis using this coefficient reveals that in most cases followers of the influential author cites the sleeping beauty paper after he cites it but the number of citations received from them are not substantial enough to explain the observations made in rank pattern analysis.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this information age, with rapid advancement in technology, while scientific literature is easily accessible to, we also have excessive amount of literature getting published in various conferences and journals in almost every field of science. For any researcher, the task of finding the best relevant articles is by no means a trivial task as not all of them are of the same quality. With information available in such excess, despite tools like Google scholar, a person neither has time nor the patience to go through all the seemingly relevant articles and choose the ones which are most appropriate for his current research endeavour. While selecting relevant articles for his research endeavour, the person ends up using personal heuristics.
Citation count of an article, reputation of its authors, the publishing journal's prestige, etc. are some of the obvious and natural heuristics that one is expected to use. The use of heuristics by people in choosing the articles to cite in their research creates the possibility that a good quality article may get ignored and gives advantage to influential authors, papers and journals of the field. But with different people using different heuristics, and exchange of ideas taking place amongst them on interaction makes task of tracking popularity of a member in a citation network a tricky but interesting problem. Given such scenario it would be very helpful to identify the members displaying similar popularity patterns and characterise the factors that play a role in rise or fall of the popularity of members following a particular popularity pattern. Being able to correctly characterise the factors that are responsible for the rise and fall of popularity can give us insights into the behavior of scientific communities.

In heterogeneous citation networks, different types of members interact and influence each other in different ways and popularity of a member at any point in his/her lifetime would depend both on itself and the members with whom his/her interaction has taken

place. Based on some commonly observed popularity patterns followed by the members, we can classify them as being **shooting star, flash in the pan, consistent performer** or **sleeping beauty**.

- **Shooting Stars :** Those who become very popular immediately after coming into existence, showing lot of promise early on but in the long run, prove to be of little value.
- **Flash In The Pan :** Those who keep gaining intense attention every now and then for short durations.
- **Consistent Performer :** Those who consistently produce useful output for the scientific community.
- **Sleeping Beauty :** These are the rare members who are ignored completely when they come into existence, but their importance is manifested much later in time and become extremely popular.

In the project, we have concentrated on papers that are rare but interesting: **Sleeping Beauty**.
We have identified sleeping beauty papers in the Pubmed and computer science subset of Microsoft Academic Network Dataset using a recently published technique.
Discrepancy in the proposed technique for identifying **Wakeup Point** of sleeping beauties, time when sleeping beauty paper starts becoming popular after sleeping for a long time, were observed. To address the issue, we instead define **Wakeup Period** and **Peaking Period** which characterize the evolution of sleeping beauty papers more aptly than a single wakeup point.
In **Rank Pattern Analysis** we observed extensive increase in sleeping-beauty's citation count around the time an influential author cites it, this motivated us to study and quantify the effect of influential authors in waking and peaking period. To account for this phenomenon we introduced a novel notion of **Followership** in heterogeneous citation network in which we provide a score, **Following Coefficient**, to quantify the followership strength of an author with respect to another author.

Chapter 2

Background

2.1 Temporal Citation Network

A citation network can be thought of as a usual graph $G = (V, E)$, except that each vertex and each edge has different properties. Moreover, the network is temporal, i.e., it changes with time, which is a crucial point in our analysis. To capture these properties, Han, et al. [8] have formalized the concept of a heterogeneous network. So, basically, we now have a network $G = (V, E, T_V, T_E, f_V, f_E, t_V, t_E)$, where V and E have their usual meaning, and T_V, T_E are type-sets for V and E. f_V, f_E are functions mapping vertices and edges to their corresponding types. The functions t_V and t_E are those specifying the year of appearance of vertices and edges respectively. In the present context, we can fix $T_V := \{P, A, V\}$, where P, A, and V denote paper, author, and venue respectively. Similarly, T_E could be fixed as $T_E := \{PP, PA, PV\}$, where PP, PA, and PV denote citation, authorship, and paper-venue respectively.
A heterogeneous citation network models the real world scenario more closely and appropriately, as it preserves semantic information in it by explicitly differentiating between different types of nodes and edges. E.g., in a heterogeneous network, if two author nodes have a common paper node as a neighbor, we know that they have co-authored that paper. But in a homogeneous network, we lose all information related to a paper, as all nodes are considered to be of the same type. So, we are only be able to say that the two nodes are co-authors, but we can't find out which paper they have co-authored.

2.2 Popularity

We see popularity as a property of entities of our citation network which varies over time. Let π_t be the popularity function, which, at a particular time t, assigns a value between 0 and 1 to the entities of our network.

2.2.1 Indicators of popularity

Citation: It is an acknowledgment that an author gives to a paper indicating that the contents of the paper were helpful for him in his research work. It indicates that the contents of the cited paper played a role in adding something new to the community. Total number of citations can only increase over time, but the rate of accumulating citations can decrease as the content of the article becomes obsolete, or increase when its content becomes relevant for the scientific community. Hence its popularity also increases or decreases with time. For instance, in the case of sleeping beauties, their citation rate is very low early in its life but increases later. On the contrary, the citation rate of shooting star is high early, but decreases at later times. So, popularity of any entity throughout its lifetime is anything but time independent. Also an entity which may be popular at one time, may become obsolete at another. Due to the "sleeping beauty" and "shooting star" effects it is difficult to predict the final number of citations based on number of citations accumulated in early years. To use citations in order to measure the popularity one must consider factors like discipline, reputation, etc. These factors have influence on citation behavior of a paper.

H-Index An author has an H-Index of h, if h of his papers have at least h citations. It represents top h papers all of which have more than or equal to h citations each. This gives information about both number of papers and number of citations. However, it ignores the actual number of citations received above h. It can be a good comparator for a discipline where citation rates are similar. But H-Index tends to be biased against the young researchers who are expected to have published a smaller number of papers. [5]

Chapter 3

Related Work and Methodologies

For a scientific article having less number of citations can become a stigma as people tend to evaluate on the basis of number of citations received. [4] Dalen et. al. performed a statistical analysis of the duration between publication and first citation. In contrast to above belief, the results showed that just because an article has not received any citation till now doesn't imply that it will never be cited in the future. Authors use survival analysis in their study to conclude this fact about seemingly dead papers [7].

Sleeping beauty is an entity whose importance in the community is not recognized for several years after coming into existence. Talking in the context of papers, it is a paper that receives very less or no citations after getting published. It stays like that for a longer duration and then shows a sudden increase in its number of citations. It would be interesting to explore the reason behind some valuable research findings receiving poor attention in early phase of their life. According to the essays published in science magazine by Eugene Garfield [2], some plausible reasons could be that:

1. The presented content is not in agreement with the current accepted theory, hence taking time for others to assimilate it.

2. It is difficult to understand and hence is avoided by most people to go through it completely.

3. Current state of the technology renders it infeasible i.e even though the idea is good but it can't be currently put to practice. [1]

According to Van Raan a sleeping beauty can be seen to have these 3 qualities [3]

1. **Length of sleep :** Duration of sleeping period
2. **Depth of sleep :** Average number of citations received during sleep period
3. **Awake intensity :** Number of citations accumulated during 4 years after sleeping period

Therefore any technique using this criteria to identify the sleeping beauty would require knowledge of some threshold values. Choosing these thresholds in itself may be a hard problem. Another method proposed by Quing Ke et. al [3] to identify sleeping beauty is a non parametric approach that takes into account the citation history of the paper. Method proposes and utilizes the concept of beauty coefficient which is calculated for the paper given its **Citation Pattern**, defined as the sequence of the number of citations received by the paper in each year since it was published upto the year corresponding to maximum number of citations. Authors attempt to use the citation pattern to analyse manner in which a paper reaches its peak. Beauty coefficient is defined as follows:-

$$\boldsymbol{B = \sum_{t_0}^{t_m} \frac{\frac{C_{t_m} - C_{t_0}}{t_m}.t + C_0 - C_t}{max\{1, C_t\}}}$$

where,

t = Age of paper in years

C_t = Citation received in t^{th} year after getting

C_t = Citation received in t^{th} year after getting publised

C_{t_m} = Citations received in its peak year

C_{t_0} = Citations received in its publishing year

$l_t = \frac{C_{t_m} - C_{t_0}}{t_m}.t + C_0$, is the line through passing through the point (t_0, C_0) and (t_m, C_{t_m}).

Beauty Coefficient is greater for the papers that exhibit sleeping beauty behavior. It also ensures that a sleeping beauty receives more citation in later years of its life, by appropriately penalizing it. Instead of classifying a paper as sleeping beauty the approach assigns a beauty coefficient to each paper, higher the beauty coefficient higher is the probability of it being a sleeping beauty.

Wakeup point t_w of sleeping beauty is defined by them as the time t at which the distance of the point (t, C_t) from line l_t is maximum.

$$\boldsymbol{t_w = arg\{max\{d_t\}\}, t \leq t_m}$$

$$\boldsymbol{d_t = \frac{|(C_{t_m} - C_0)t - t_m C_t + t_m C_0|}{\sqrt{(C_{t_m} - C_{t_0})^2 + (t_m)^2}}}$$

3.1 Our Work

We identified sleeping beauty papers using beauty coefficient mentioned earlier. To address the issues in identification of wakeup point we argue against its importance in the evolution of sleeping beauty paper and define two new concepts of waking period and peaking period. Our findings in rank pattern analysis motivated us to study the role of influential authors in both these periods. To account for the findings of rank pattern analysis we studied average rank of influential authors' papers' references for duration 1990-2015 and performed follower pattern analysis based on our following coefficient.

3.1.1 Waking Period, Peaking Period and Wakeup Point

As sleeping beauty papers gain recognition after being ignored for a long time, it is very intriguing to find and study the factors that ensue such drastic change in its course of evolution. A precise identification and robust understanding of such factors can help us revive more of such ignored but useful papers.

Quing Ke et. al attempt to do so by defining **wakeup point** of sleeping beauty as the year in which abrupt change in its accumulation of citation occurs. We found that their proposed method for its identification, in many cases, tends to report the year with no citations as the wakeup point. This defeats the very purpose of finding it, as there is nothing to be analysed in such cases.

Although citation pattern of sleeping beauty makes it very tempting to look for a wakeup point, as defined in the early literature related to it. On the contrary a careful observation of its citation pattern would indicate that this delayed accumulation of citations builds up gradually over a period rather than being a sudden event that can be attributed to a particular point. Hence, rather than focusing on finding a point, we chose to partition its late citation accumulation period into waking period and peaking period as defined below.

Waking Period : We define waking period as a closed discrete interval $P_{wake} = [t_s, t_e]$, where t_s is taken as input from user based on the inspection of sleeping beauty's citation pattern graph and t_e is such that :

$C_{sleep} = \sum_{t \in P_{sleep}} C_t$, $P_{sleep} = [t_0, t_s - 1]$

$C_{late} = \sum_{t \in P_{late}} C_t$, $P_{late} = [t_s, t_m]$

$C_{wake} = \sum_{t \in P_{wake}} C_t$, $P_{wake} = [t_s, t_e]$

$R_{wake} = \frac{C_{wake}}{t_e - t_s + 1}$, $R_{sleep} = \frac{C_{sleep}}{t_s - t_0}$

$R_{wake} > R_{sleep}$, $C_{wake} > C_{sleep}$

and $C_{wake} \geq 0.1 * C_{late}$

Peaking Period : Once we get t_e for waking period, peaking period of sleeping beauty can simply be defined as a closed interval $P_{peak} = [t_e + 1, t_m]$, where t_m is the year in which it receives maximum number of citations.
Observing sleeping beauty's citation pattern graph indicates that during sleeping period of its evolution, it either gathers no citations or very few citations in an intermittent manner spread over a long period of time. Later on, however, its citation accumulation behavior shows a slow but steady increase for a short while before it finally speeds up towards its maximum. Hence, partitioning late citation accumulation phase of sleeping beauty into two periods, with wakeup period corresponding to slow-steady increase and peaking period corresponding to speeding towards its maximum, captures the evolutionary changes better than just a single wakeup point. Observation made in rank pattern analysis, described in section 4.4, motivated us to study and quantify the effect of authors' rank in both the periods, which lead to useful insight in our analysis.

3.1.2 Role Of Influential Authors In Evolution Of Sleeping Beauty

Influential authors play an important role in the evolution of science. Their research work acts as foundation for future research as an important literature assisting other authors in their research work. Therefore we were tempted to find out whether they play any role in evolution of sleeping beauty. For this we performed Rank Pattern Analysis using page rank as the measure of author's influence. In this analysis we looked at the rank of authors who cited

sleeping beauty in its course of evolution. Analysis revealed that the sleeping beauty begins to accumulate citations at an intense rate around the time an influential author cites it, which had otherwise struggled to gain any citation for a long period.
we also looked at the average rank of the references made by the authors during 1990-2015, a considerable difference is observed in the average rank of high ranked authors' references and low ranked authors' references.
To account for the effect of influential authors on citation accumulation of sleeping beauty, we performed Follower Pattern Analysis based on followership in heterogeneous citation network which we described in next subsection 3.1.3.

3.1.3 Followership In Heterogeneous Citation Network

In this section, we give a general approach to quantify followership between two authors in heterogeneous citation network. We describe it in terms of abstract functions that aim to quantify certain aspect of interaction between two authors, assumed to have a role in developing followership behavior of one author with respect to another.

Scientific community, at any point of time has certain fraction of authors that are influential and their ideas/research-output get more recognition as compared to others'. Their influence and perceived authority is helpful in spreading their ideas faster. A new idea proposed by them is likely to face less resistance in getting adopted and becoming prevalent among other authors in citation network. However, ideas of a particular influential author may not get noticed and adopted by every other author in citation network, but there may exist some authors who do refer to their work very frequently and extensively. We identify such authors as his followers and aim to develop a function **Followership(A,B)**, which gives us a score, **Following Coefficient**, that quantifies the followership strength of Author-A with respect to Author-B. In doing so, we make some assumption about the general behavior of authors based on their influence in scientific community. We define function $\boldsymbol{\phi(X)}$ as the influence of author $\boldsymbol{X}$ in citation network, attributes such as page rank, node centrality etc. can be used for this purpose. This helps us to separate influential authors from non influential authors based on a particular threshold, resulting in potential followers and leaders in citation network.
We break down calculation of followership strength of Author-A with respect to Author-B in 3 parts as follows :

1. **Defining Following Affinity :** We define affinity function, $\boldsymbol{0 \leq \alpha(\phi(A), \phi(B)) \leq 1}$, which quantifies the tendency of Author-A to follow Author-B based on following Assumptions:

(a) **Followership is not an exclusive relationship:** An author can be a follower of more than one author.

(b) **Followership is not a transitive relation :** If A is a follower of B and B is a follower of C then A is not necessarily a follower of C.

(c) **Followership is not a symmetric relation :** If A is a follower of B and B is not necessarily a follower of A.

(d) **Affinity of Author-A with respect to Author-B depends on influential status of A and B according to Table 3.1.**

$\boldsymbol{\phi(A)}$	$\boldsymbol{\phi(B)}$	$\alpha(\phi(A), \phi(B))$
High	High	Very low (ego/competition)
Low	High	$\propto \phi(B)$
Low	Low	0

Table 3.1: **Affinity Between Authors**

2. **Defining and Quantifying behavior of follower :** We need to define and quantify the strength of the behavior a follower is expected to display in its interaction with the followee. We define function $\boldsymbol{\beta_i(A, B)}$ which quantifies the strength of Author-A's i^{th} behavior with respect to Author-B.

3. **Followership Score :** It is the sum of behavior strength weighted by affinity, where $\boldsymbol{N}$ is the number of behaviors considered.

$$\sum_{i=1}^{N} \boldsymbol{\beta_i(A, B) * \alpha(\phi(A), \phi(B))}$$

3.1.4 Following Coefficient

We calculate following coefficient to quantify followership strength of one author with respect to another author using the approach described in section 3.1.3. Functions required by the approach for its calculation are described below.

1. **Influence Function $\boldsymbol{\phi(X)}$:** We use rank, $\boldsymbol{R(X)}$, of the author $\boldsymbol{X}$ based on its page rank [6] score as a measure of his/her influence.

2. **Affinity Function $\boldsymbol{\alpha(R(A), R(B))}$:** We choose a threshold **r**, based on which our Affinity function is as follows:

(a) $\alpha(\mathcal{R}(\mathcal{A}), \mathcal{R}(\mathcal{B})) = 0$ if $\mathcal{R}(\mathcal{A}) \geq \mathbf{r}$ and $\mathcal{R}(\mathcal{B}) < \mathbf{r}$

(b) $\alpha(\mathcal{R}(\mathcal{A}), \mathcal{R}(\mathcal{B})) = 0$ if $\mathcal{R}(\mathcal{A}) < \mathbf{r}$ and $\mathcal{R}(\mathcal{B}) < \mathbf{r}$

(c) $\alpha(\mathcal{R}(\mathcal{A}), \mathcal{R}(\mathcal{B})) = \frac{\mathcal{R}(\mathcal{B})}{\mathcal{N}}$ if $\mathcal{R}(\mathcal{A}) < \mathbf{r}$ and $\mathcal{R}(\mathcal{B}) \geq \mathbf{r}$,
$\mathcal{N}$ denotes total number of authors in dataset.

(d) $\alpha(\mathcal{R}(\mathcal{A}), \mathcal{R}(\mathcal{B})) = e^{-\frac{\mathbf{r}}{|\mathcal{R}(A) - \mathcal{R}(B)|}}$ if $\mathcal{R}(\mathcal{A}) \geq \mathbf{r}$ and $\mathcal{R}(\mathcal{B}) \geq \mathbf{r}$

We define affinity function between two high rank authors as **(d)**. We assume that ego and competition among them is inversely proportional to the closeness of their rank. Therefore, closer is their rank lesser is the tendency of following each other.

3. **Behavior Function $\boldsymbol{\beta}$:** We assume that if an author follows another other author, he tends to cite papers cited by him after small duration of him citing it. Author citing an otherwise ignored paper within small duration of some other author citing it, insinuates strong citation copying behavior of his with respect to that author.

   Let, $\boldsymbol{\rho_a}$ be the set of references of author $\boldsymbol{a}$, $\boldsymbol{\tau_a(r)}$ be the sorted list of years in which author $\boldsymbol{a}$ cited $\boldsymbol{r}$, $\boldsymbol{\omega(a,b)} = \boldsymbol{\rho_a} \cap \boldsymbol{\rho_b}$. To calculate behavior strength $\boldsymbol{\beta(a,b)}$, for each reference $\boldsymbol{r} \in \boldsymbol{\omega(a,b)}$ we calculate $\boldsymbol{\tau_b(r)}$ and $\boldsymbol{\tau_b(r)}$. For each $\boldsymbol{t_a} \in \boldsymbol{\tau_a(r)}$ we define $\boldsymbol{F_{t_a}}$ in two ways : $\boldsymbol{F_{t_a}} = \frac{\boldsymbol{T_1}}{\boldsymbol{T_2}}$ or $\boldsymbol{F_{t_a}} = \boldsymbol{e}^{\frac{\boldsymbol{T_1}}{\boldsymbol{T_2}}}$ where

   $\boldsymbol{T_1} = \boldsymbol{t_a} - \boldsymbol{t_b}$ and $\boldsymbol{T_2} = \boldsymbol{t_a} - \boldsymbol{t_{<a>}}$ such that:

   - $\boldsymbol{t_b}$ is the smallest number $\boldsymbol{t} \in \boldsymbol{\tau_b(r)}$ such that $\boldsymbol{t_a} - \boldsymbol{t} \leq \boldsymbol{\delta}$.
   - If $\boldsymbol{t_a}$ is the smallest number in $\boldsymbol{\tau_a(r)}$ then $\boldsymbol{t_{<a>}}$ is the year of $\boldsymbol{a's}$ first publication otherwise it is the largest number in $\boldsymbol{\tau_a(r)}$ less than $\boldsymbol{t_a}$.

   Behavior score $\boldsymbol{\beta(a,b)}$ is calculated as follows :

$$\sum_{\boldsymbol{r} \in \boldsymbol{\omega(a,b)}} \sum_{\boldsymbol{t_a} \in \boldsymbol{\tau_a(r)}} \boldsymbol{F_{t_a}} * \boldsymbol{\Gamma_{t_a}(a,b)} * \boldsymbol{\Lambda_{t_a}(r)}$$

   $\boldsymbol{F_{t_a}}$ measures that how immediately an author cited a paper, which was otherwise ignored by him for long time, after his followee cited it.

   $\boldsymbol{\Gamma_{t_a}(a,b)}$ is the probability of $\boldsymbol{a}$ copying the citation of $\boldsymbol{b}$ in year $\boldsymbol{t_a}$, as defined in section 3.1.5.

   $\boldsymbol{\Lambda_{t_a}(r)}$ is the percentage of papers with rank greater than rank of $\boldsymbol{r}$ in year $\boldsymbol{t_a}$. Its higher value would indicate that an author has cited an unpopular paper after it is cited by his followee, which indicates a strong followership behavior.

3.1.5 Citation Copying Probability, $\Gamma_t(a, b)$

We want the probability that an author has cited a paper because some other author whom he follows has cited it. We take two approaches to model the citation copying probability. ***Let***, $\boldsymbol{\rho_a(t)}$ and $\boldsymbol{\rho_b(t)}$ be the set of references of author $\boldsymbol{a}$ and $\boldsymbol{b}$ upto year $\boldsymbol{t}$, respectively. $\boldsymbol{P_a(t)}$ and $\boldsymbol{P_b(t)}$ be the set of author $\boldsymbol{a's}$ and $\boldsymbol{b's}$ paper upto year $\boldsymbol{t}$, respectively.

Approach $\mathbb{I}$: In this approach we say that author $\boldsymbol{a}$ first reads author $\boldsymbol{b's}$ paper and then cites one of the references in those papers. Probability of author $\boldsymbol{a}$ reading author $\boldsymbol{b's}$ paper in year $\boldsymbol{t}$ is $\frac{|\rho_a(t) \cap P_b(t)|}{|\rho_a(t)|}$. Probability of author $\boldsymbol{a}$ citing references of author $\boldsymbol{b's}$ paper in year $\boldsymbol{t}$ is $\frac{|\rho_a(t) \cap \rho_b(t)|}{|\rho_b(t)|}$.

Required probability is simply the product of above two probabilites :

$$\Gamma_t(a, b)^{\mathbb{I}} = \frac{|\rho_a(t) \cap P_b(t)|}{|\rho_a(t)|} \times \frac{|\rho_a(t) \cap \rho_b(t)|}{|\rho_b(t)|}$$

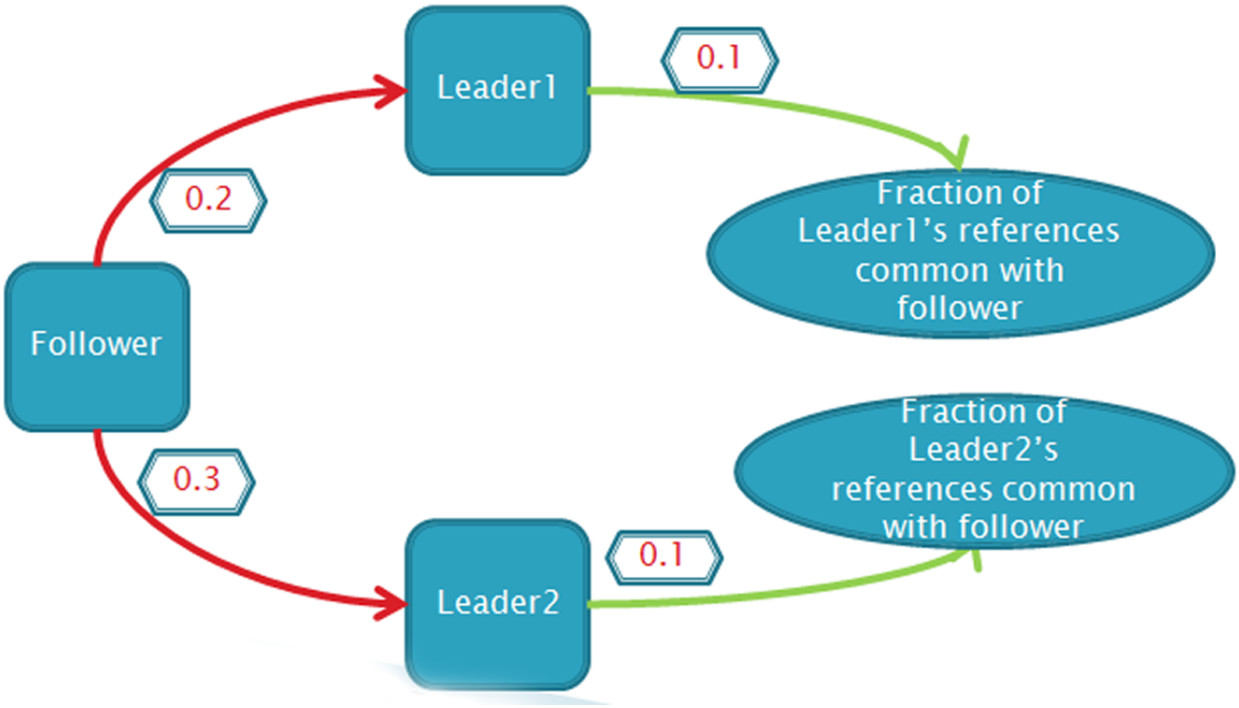

Figure 3.1: Citation Copying Probability Approach I

Approach $\mathbb{III}$: In this approach instead of considering all the common references of author $\boldsymbol{a}$ and $\boldsymbol{b}$ as an indication of author $\boldsymbol{a}$ copying $\boldsymbol{b's}$ citations, we only consider those citations that are in author $\boldsymbol{b's}$ paper that author $\boldsymbol{a}$ has actually read.
Let, $\boldsymbol{P_{ab}(t) = \rho_a(t) \cap P_b(t)}$, $\boldsymbol{R_{ab}(t)}$ be the set of references of papers in $\boldsymbol{P_{ab}(t)}$ and $\boldsymbol{\rho_{ab}(t) = R_{ab}(t) \cap \rho_a(t) \cap \rho_b(t)}$. Required probability is :

$$\boldsymbol{\Gamma_t(a,b)^{\mathbb{III}} = \frac{|\rho_{ab}|}{|\rho_a|}}$$

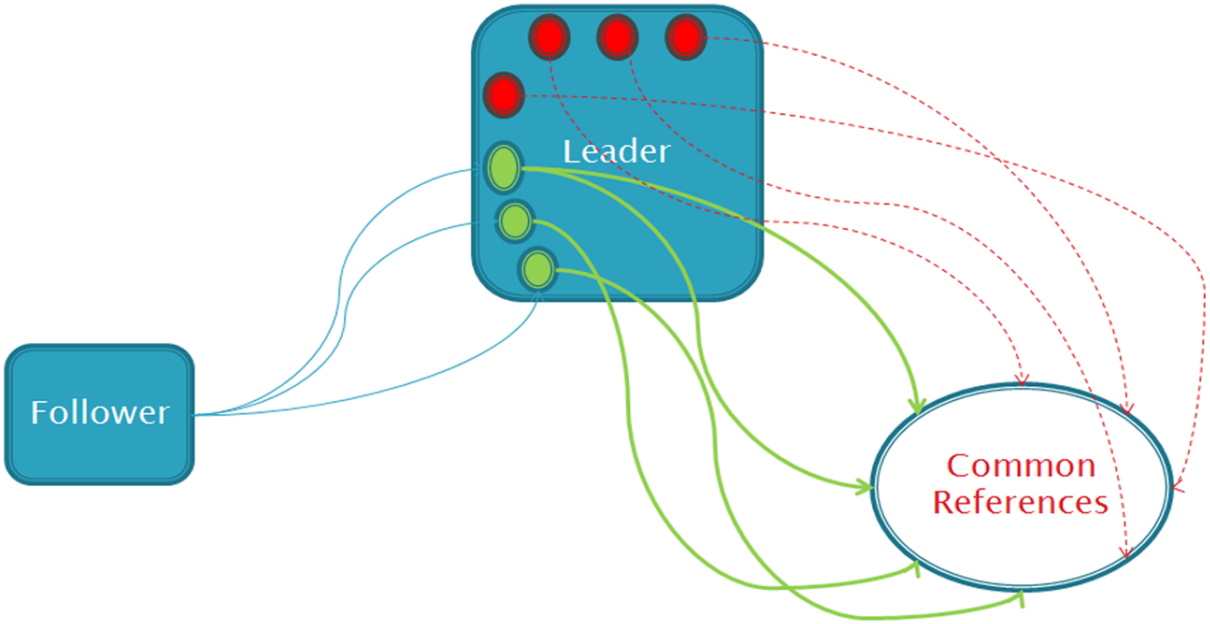

Figure 3.2: Citation Copying Probability Approach II

Chapter 4

Experiments and Results

For our experiments we have considered following **dataset**

- **Pubmed Dataset** which contains journal citations and abstracts for research papers published in biomedical field containing 99214 papers, 756613 references and 212312 authors.
- **Computer Science Subset Of Microsoft Academic Network(CS-MAN)** which contains research papers published in computer science field containing 528719 papers, 1061231 references and 776560 authors.

Both datasets have information about the name of the authors and the publishing year for each paper.

4.1 Identifying Sleeping Beauties

- **Input :** Citation Pattern
- **Output :** Beauty Coefficient

For calculation of beauty coefficients of papers in the dataset we need citation pattern of the paper which is the number of citations that the paper received in each year since it was published up to the year in which it received maximum number of yearly citations. Citation pattern can be easily extracted from the file containing references and the file containing publishing year of each paper. Beauty coefficients of sleeping beauties identified in both datasets are listed in Table 4.1 and 4.2.

Paper Id	Title	Beauty Coefficient
938	Junctional complexes in various epithelia	184.79
4833	Voltage oscillations in the barnacle giant muscle fiber	128.02
3385	Branch input resistance and steady attenuation	125.81
4055	Electrophysiology of guinea-pig cerebellar nuclear cells	99.92
454	physiological mechanism for Hebb's postulate of learning	99.71

Table 4.1: Top five sleeping beauties in Pubmed.

Paper Id	Title	Beauty Coefficient
29348	Low Density Parity Check Codes	747.87
412928	Particle swarm optimization	389.80
344538	Identity-based cryptosystems and signature schemes	306.61
523387	An algorithm for tracking multiple targets cells	188.38
235719	The complexity of theorem-proving procedures	176.10

Table 4.2: Top five sleeping beauties in CS-MAN.

4.2 Experiments Addressing Wakeup Point Discrepancies

The method described in for identifying wakeup point reports year with no citations as wakeup point for many sleeping beauties in both datasets as shown in Table 4.3 and 4.4. For addressing the discrepancies in the proposed method for identifying wake up point we performed two experiments.

4.2.1 Wakeup Point Experiment 1

We defined it as that time in the lifetime of paper when its **right citation rate** becomes greater than its **left citation rate**, because given the fact that the paper is sleeping beauty we know that it has more citations in later point in its life when it started to peak.

At $\boldsymbol{i^{th}}$ time step in its lifetime

Left Citation Rate = $\frac{1}{t_i}\sum_{j=0}^{i-1} C_{t_j}$

Right Citation Rate = $\frac{1}{t_m - t_i}\sum_{j=i}^{m} C_{t_j}$

Results of this experiments were not as expected and failed to address the issues pointed out in the proposed method.

Paper Id	Wakeup Point	Title	C_{wake}
520628	2002	Fuzzy clustering with a fuzzy covariance matrix	0
520119	1998	Stochastic approximation algorithms and applications	0
29348	2001	Low Density Parity Check Codes	5
523387	2000	An algorithm for tracking multiple targets	0
344538	2004	Identity-based cryptosystems and signature schemes	6
412928	2004	Particle swarm optimization	16
466547	2004	$n^5/2$ Algorithm for Max. Matchings in Bipartite Graphs	0

Table 4.3: Wakeup point of sleeping beauties in CS-MAN and citations received in wakeup year

Paper Id	Wakeup Point	Title	C_{wake}
74	2005	Fuzzy clustering with a fuzzy covariance matrix	0
4833	2009	Voltage oscillations in the barnacle giant muscle fiber	0
4055	2006	Electrophysiology of guinea-pig cerebellar nuclear cells	0
3385	2010	Branch input resistance and steady attenuation	0
938	2010	Junctional complexes in various epithelia	3

Table 4.4: Wakeup point of sleeping beauties in Pubmed and citations received in wakeup year

4.2.2 Wakeup Point Experiment 2

For this experiment we slightly modified the expression for **left citation rate** and **right citation rate** as follows :

At $\boldsymbol{i^{th}}$ time step in its lifetime

Left Citation Rate = $\frac{1}{t_i + C_{total}} \sum_{j=0}^{i-1} C_{t_j}$

Right Citation Rate = $\frac{1}{t_m - t_i + C_{total}} \sum_{j=i}^{m} C_{t_j}$

Although experiment results in the kind of behavior that we were hoping to get in Experiment 1, we do not have any mathematical explanation to establish a strong logical connection between the behavior exhibited and wakeup point of sleeping beauty.

4.3 Waking and Peaking Period

We have highlighted the limitation of wakeup point and utility of Waking and Peaking period in capturing evolution of sleeping beauty papers in section 3.1.1.

- **Input :** Citation Pattern, Starting point of waking period, chosen by user based on visual inspection of the graph of citation pattern of sleeping beauty
- **Output :** Waking Period, Peaking Period, Citations Accumulated and Citation Accumulation Rate of sleeping beauty in these periods, listed in Table 4.5 and 4.6.

ID	P_{wake}	C_{wake}	R_{wake}
29348	1999-2003	42	8.4
412928	2002-2006	192	38.4
344538	1999-2005	51	7.29
523387	2001-2003	14	4.67
235719	1991-2007	140	20.7
466535	1998-2001	24	6
34678	1999-2006	60	7.5

Table 4.5: Waking Period Statistics Of CS-MAN

ID	P_{peak}	C_{peak}	R_{peak}
29348	2004-2009	243	40.6
412928	2007-2009	495	165
344538	2006-2007	96	48
523387	2004-2009	86	14.33
235719	1998-2008	737	67
466535	2002-2008	118	16.85
34678	2007-2008	44	22

Table 4.6: Peaking Period Statistics Of CS-MAN Dataset

4.4 Rank Pattern Analysis

Objective of this experiment was to study the role of influential authors in waking up sleeping beauties. It requires three steps :-

1. **Identify sleeping beauties in the dataset:** We identified sleeping beauties using the method proposed in [3].
2. **For each year in citation history in which the sleeping beauty received citations, calculate rank of the authors that cited the paper in that year :-**

We first construct the subset of the references that were published upto and before the year for which we need the rank and then construct the citation graph on this subset and give it as an input to page rank algorithm to calculate the page rank of papers.

To calculate the rank of the author we sum the rank of the papers published by them in this subset and use this sum value to find the rank of authors.

3. **Plots :** We are plotting the reciprocal of the highest rank of the authors for each year in the citation history of sleeping beauty in which it received non zero citations. Resulting plots can be seen in figure 4.1 and 4.2.

Note:- We are plotting reciprocal because author with higher influence will have lower rank index.

4.5 Average Rank Of Authors' References

- **Input :** Rank of authors and papers based on page rank score for each year in the duration 1990-2015
- **Output :** A plot of **Year Vs Average Rank** of references of group of authors, where authors are grouped on the basis of their rank in that year.

For each year we grouped authors based on their rank and for each group we calculated average rank of authors' references, with better papers having lower rank index. Resulting plot is in figure 4.3.
For the purpose of this analysis we had precomputed the ranks of authors and papers for each year in the duration 1990-2015.

4.6 Influential Authors In Waking And Peaking Period

- **Input :** Citation Pattern, Rank Pattern, P_{wake}, R_{wake}, P_{peak}, C_{peak}, R_{peak}
- **Output :** Correlation of length of P_{wake}, R_{wake}, C_{peak}, R_{peak} with highest rank among citing authors in initial 3 years of the period

Observation made in Rank Pattern Analysis motivated us to study the effect of citing authors' rank in waking and peaking period of sleeping beauty. Based on the resulting correlation values, listed in Table 4.7, we can make following conclusions :

- Low correlation of authors' rank with length of waking period indicates that it doesn't help in shortening its waking period, otherwise we would have observed large negative correlation

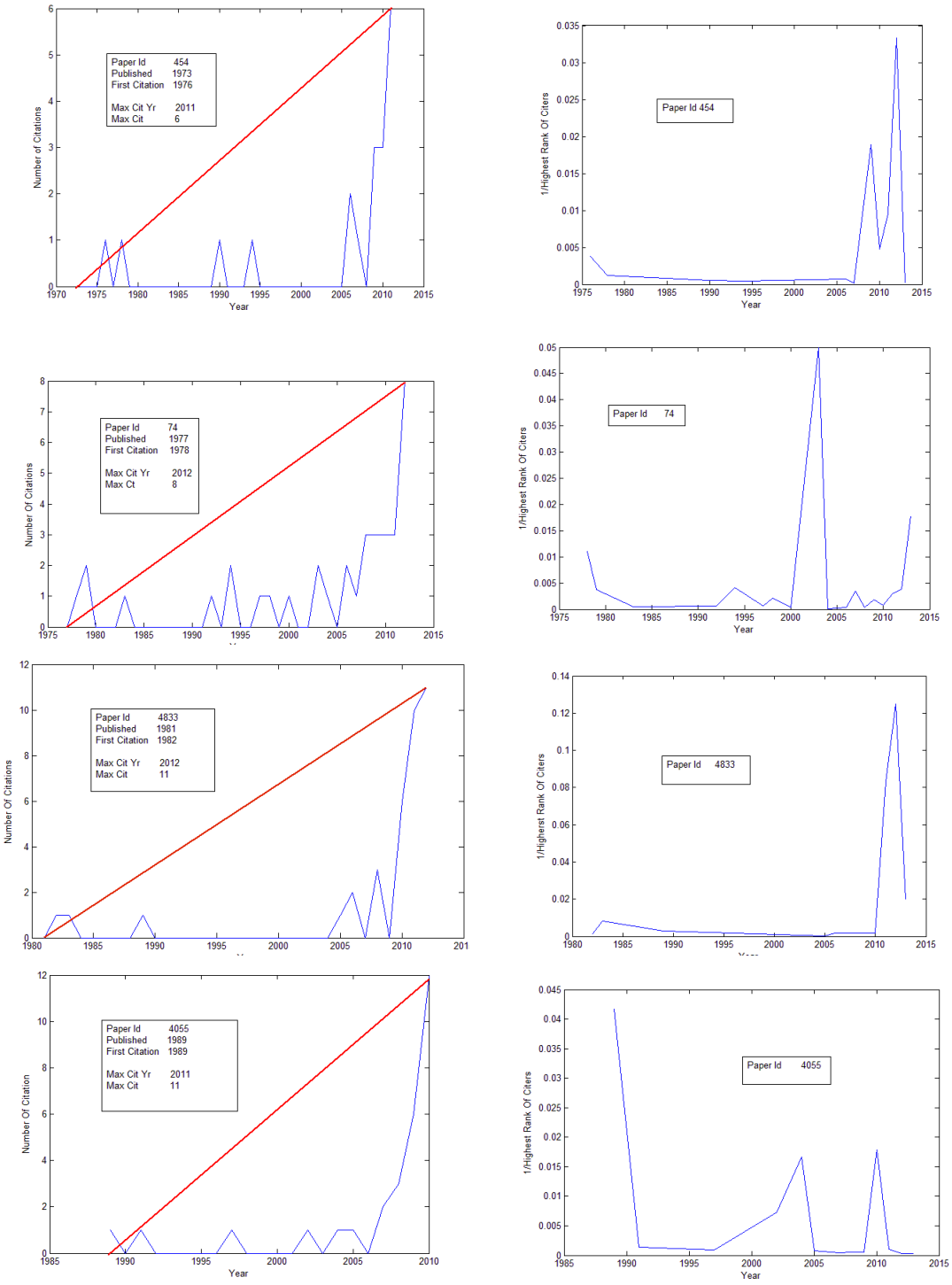

Figure 4.1: Rank Pattern Analysis For Pubmed Dataset

- Low correlation of authors' rank with citation accumulation rate in waking and peaking period indicates that it doesn't assist in gaining citations quickly.

- High correlation of authors' rank with citations accumulated in peaking period indicates that after waking up better the citing author's rank more is the citation count in peaking period.

a	$Corr(a, \frac{1}{rank})$
C_{peak}	0.76
R_{peak}	0.22
R_{wake}	0.26
length of P_{wake}	0.15

Table 4.7: Correlation values of waking and peaking period statistics with authors' rank

4.7 Follower Pattern Analysis

- **Input :** Year wise rank of authors and papers, Following coefficient between authors.
- **Output :** Graphs with following plots :
  - Year Vs Highest Following Coefficient of authors citing sleeping beauty in that year, with respect to an influential author who has cited sleeping beauty.
  - Year Vs Sleeping Beauty's yearly citation and Year Vs Sleeping Beauty's yearly citation received from authors having positive following coefficient with respect to an influential author, plotted together.

For this analysis we have used following coefficient as described earlier. For the purpose of our experiments we have arbitrarily chosen $\boldsymbol{\delta = 5}$. We have considered two choices for $\boldsymbol{F_{t_a}}$ and $\boldsymbol{\Gamma_{t_a}(a, b)}$, so we have four following coefficients for a given sleeping beauty and influential author.
As the purpose of this experiment was to explain the observation about sleeping beauty papers made in rank pattern analysis, we have excluded them while calculating following coefficient, and we have also excluded highly cited papers because citation given to them doesn't indicate citation copying behavior. Resulting graphs can be seen in figure 4.4, 4.5 and 4.6.

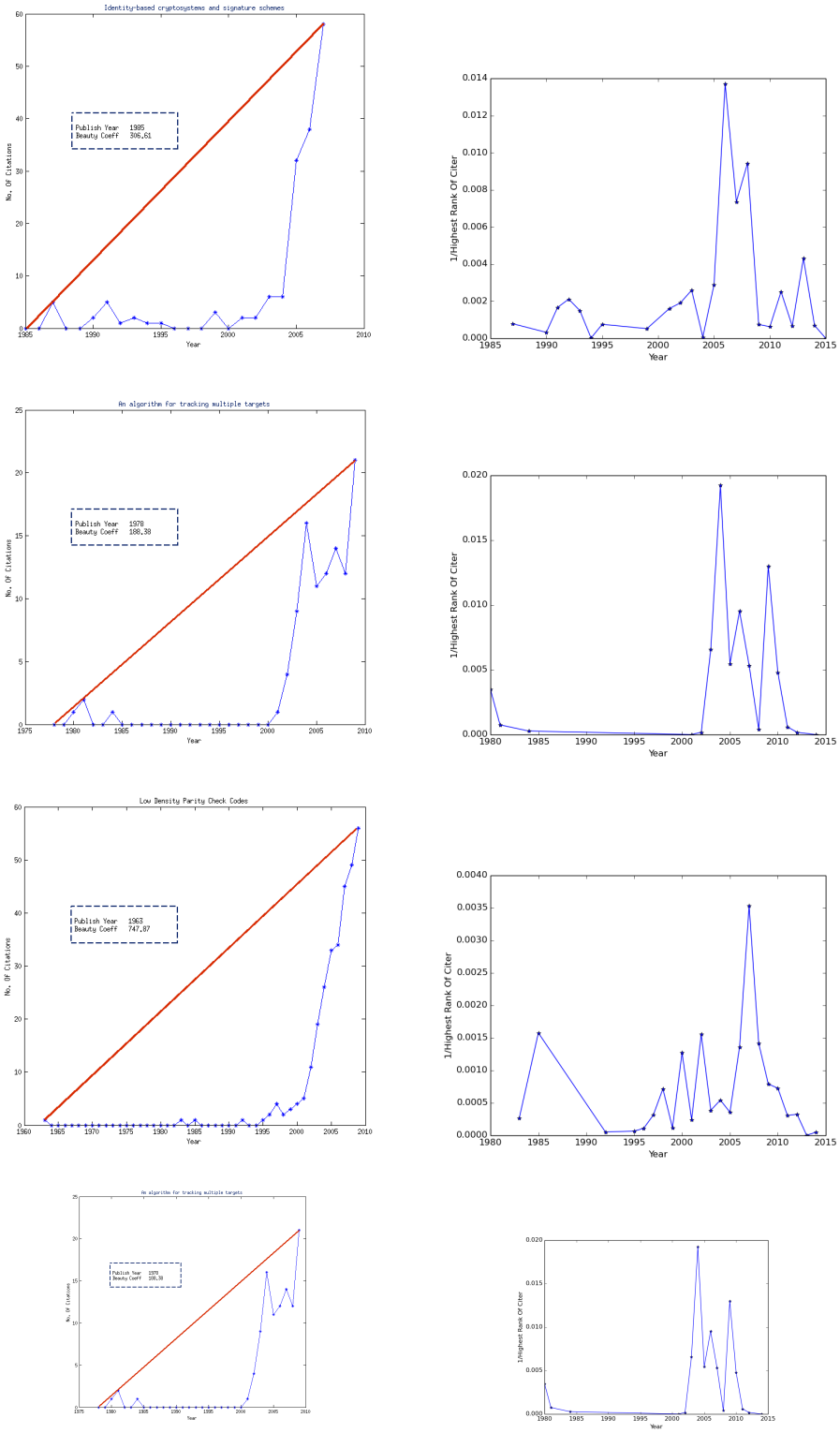

Figure 4.2: Rank Pattern Analysis For CS-MAN Dataset

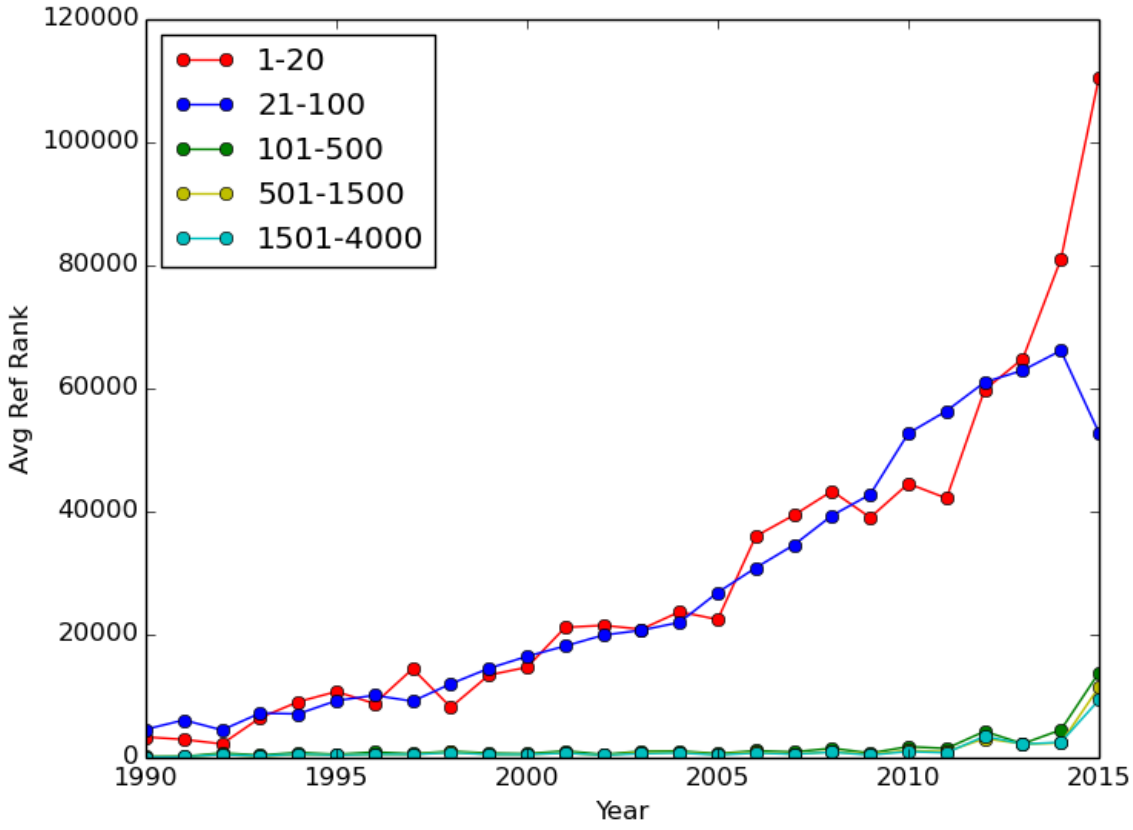

Figure 4.3: Average rank of references of authors

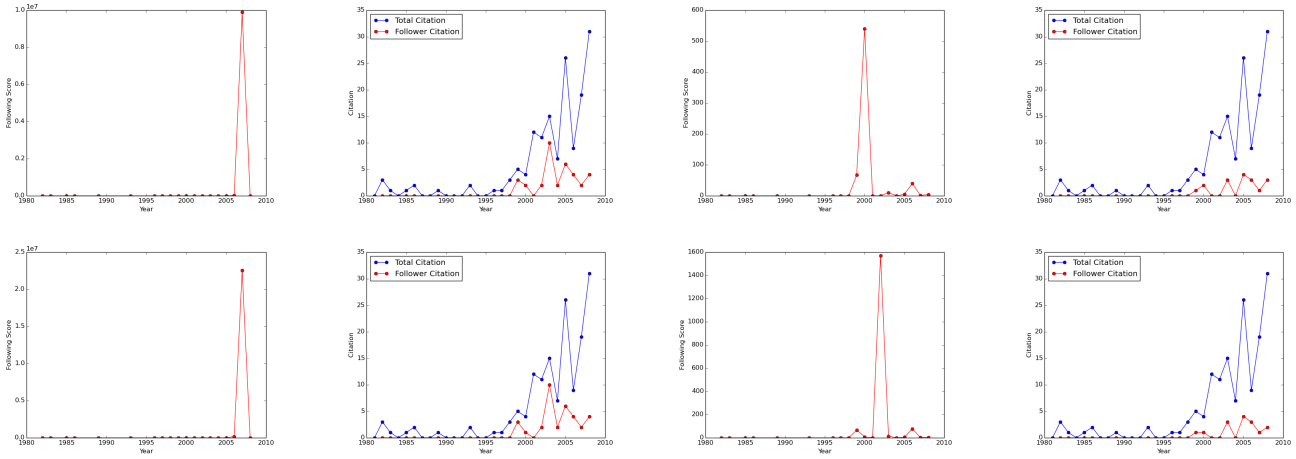

Figure 4.4: Follower Pattern Analysis Paper 466535
Author ID : 691035
Author Rank : 68
First Cited 466535 In Year 2001

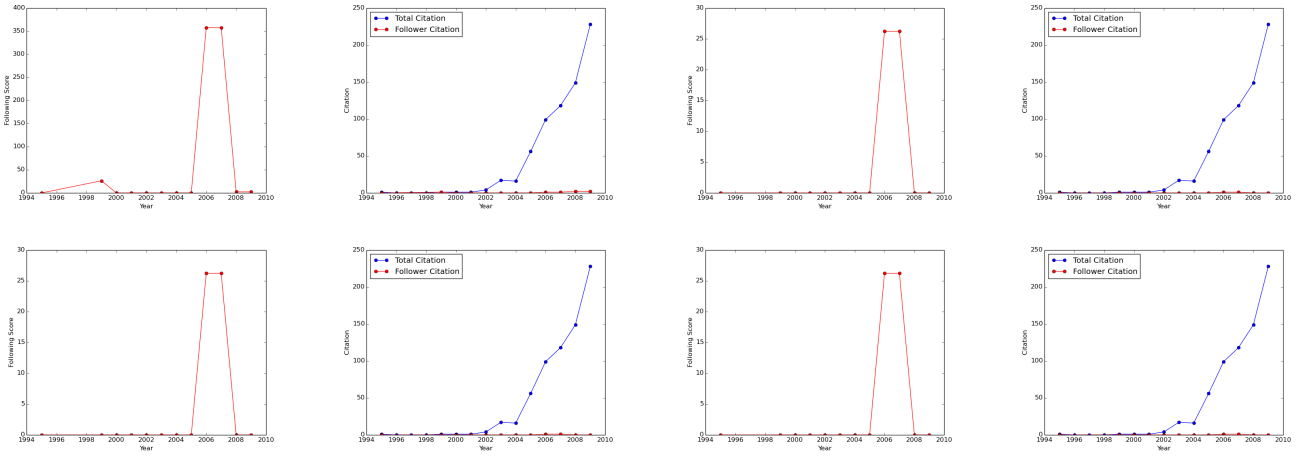

Figure 4.5: Follower Pattern Analysis Paper 412928
Author ID : 448438
Author Rank : 660
First Cited 412928 In Year 2005

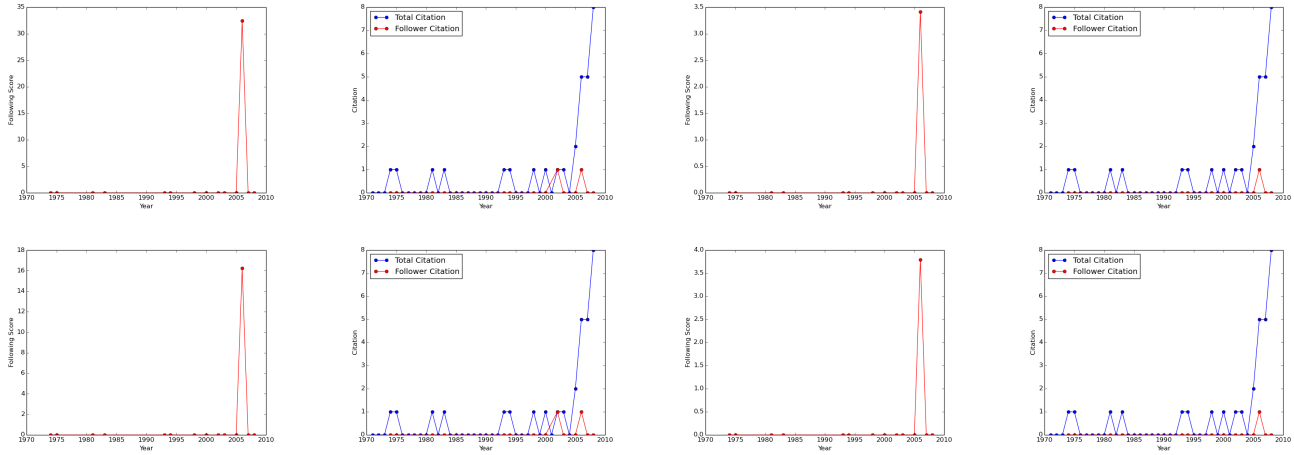

Figure 4.6: Follower Pattern Analysis for Paper 466547
Author ID : 46198
Author Rank : 52
First Cited 466547 In Year 2005

Chapter 5

Discussion

5.1 Wakeup Point Experiment

Experiment 1: In this experiment we were expecting that the graph of **right citation rate** would go above graph of **left citation rate** after starting off below it and the point at which this happens can be uniquely identified as the wake up point of sleeping beauty, but it didn't happen so.

Experiment 2: In this experiment we slightly modified the expression for citation rate that we are studying in Experiment 1, motivation was to do something similar to Laplace Correction.

Although results were as expected but we fail to explain it mathematically because unit of the quantity is not making any sense and can not be claimed to be measuring anything about the sleeping beauty under consideration.
Nevertheless, the results produced are still interesting, as for as the papers with high beauty coefficient value the two graphs are coming close to each other near the point where there is a **noticeable increase in yearly citations received by sleeping beauty after a long sleep**, and this is exactly the kind of point that we hunt for when we talk about the wakeup point of a sleeping beauty.

5.2 Rank Pattern Analysis

Interpretation of the results obtained for this analysis is very interesting. We juxtaposed two graphs for comparison:-

1. **Citation Pattern**

2. **Highest Rank Pattern**

For the top sleeping beauties what one can observe is that :

1. In the initial years after being published, only the low ranked authors are citing it and for a long time we don't see any noticeable increase in the yearly citation.

2. The time when one observes a strong spike in Highest Rank Pattern graph, meaning that some highly ranked author has cited it, within a short duration after it we can observe an increase in the yearly citations received by the paper, but this kind of change wasn't observed earlier when low rank authors were citing it in its early years.

This trend can be observed among top sleeping beauties, indicating that influential authors do play role in waking up sleeping beauties.

Table 5.1 and 5.2 list the correlation value between the highest [1/rank] ranked authors citing the paper during its citation history and the yearly citations received by the paper.

Paper ID	**Correlation**
938	0.7291
4833	0.8142
3385	0.7958
4055	-0.0053
454	0.6024
74	0.0308

Table 5.1: Correlation of citing authors' highest rank and yearly citation.(Pubmed Dataset)

Paper ID	**Correlation**
29348	0.4751
412928	0.4806
344538	0.6695
523387	0.6286
34678	0.5583
520628	0.5301

Table 5.2: Correlation of citing authors' highest rank and yearly citations.(CS-MAN)

We can see that except for the paper 4055 which has a negative correlation and paper 74 which has a very low positive correlation 0.0308, others are showing quite a strong positive correlation.
Note :- We are taking 1/Rank for the same reason as mentioned in section 4.4.

5.3 Average Rank Of Authors' References

We performed this experiment to find out if there is a difference in the reference pattern of the authors based on their ranks. In the resulting plot, figure 4.3, we can clearly observe that:

- **Average rank of influential authors' references is higher as compared to non influential authors'.**
- **With time, gap between the average rank of influential authors' references and non influential authors' references has increased monotonically.**

This suggests that in their research work influential authors also tend to consider less popular literature whereas non influential authors tend to refer only standard popular literature.
It also indicates that one performs better in his scholarly career if they take a holistic approach in their research rather than just focusing on standard and popular literature.
Sleeping beauty paper remains unpopular for a long period in its lifetime. This analysis suggests that it is more likely to be woken up by influential authors as they tend look beyond the popular literature whereas non influential ones are likely to ignore it as they mostly refer to popular literature.

5.4 Follower Pattern Analysis

As observed in Rank Pattern Analysis, citation accumulation of sleeping beauty increases in the vicinity of getting citation from a high ranked author. We conjectured that this rise is due its increased visibility amongst the followers of that highly ranked author. Sleeping beauty gets the benefit of visibility of the papers of that author amongst its followers.
For this analysis we used following coefficient described in section 3.1.4 to quantify the followership strength of one author with respect to another author. We expected that highest following coefficient of the authors who have cited sleeping beauty paper with respect to some high ranked author, would show a considerable increase only after the year that high ranked author cited it, and the citations received from them would be considerable.
Our conjecture did not hold true for the sleeping beauties identified in CS-MAN. In some cases coefficient is high even before the year when a high ranked author cited it, in some cases magnitude of highest following coefficient is negligible and in some cases the citations from those

authors is not significant as compared to the total citations received. For some cases, after staying at zero for long time, a sudden rise in following coefficient is observed just after the year when a high ranked author has cited it.
Based on the results obtained we can not conclude that the observed rise in citations of sleeping beauty is due to its increased visibility amongst the followers of some high ranked author.

Chapter 6

Conclusion And Future Work

A sleeping beauty is any real world entity whose importance is manifested in an unprecedented manner, in the social environment in which it lives, much later after coming into existence. The reason for such phenomenon will depend greatly on the behavior of other members of that society. Hence, evolution of such an entity can't be studied in isolation. In our project we aimed at studying the evolution of research papers published in heterogeneous citation network that manifest such behavior. We identified sleeping beauty papers, in the datasets mentioned, using a recently published technique. As the social setting here involves individuals engaged in scientific research competing against each other, to do better in terms of various accepted performance measures. In such a scenario a researcher is expected to be inclined towards having popular literature associated with influential/well-known author as the foundation of his research work. Aligned with this thought process and general social behavior of following celebrities motivated us to investigate the role of influential authors in the sudden attention gained by a long ignored paper.

We performed rank pattern analysis which revealed that in most cases yearly citations of sleeping beauty increases extensively around the time when it is cited by influential author. To explain the delayed recognition experienced by sleeping beauty we defined waking and peaking period and studied the effect of rank of authors who are citing it in those period. Our experiments do not indicate any role of authors' rank in waking period but in peaking period we observed high positive correlation between citing authors' rank and citations accumulated. We also studied the average rank of authors' references over a period of 1990-2015 which showed a significant difference between the average ranks of influential authors' references and non influential authors' references, indicating that influential authors also tend to consider less popular papers in their research work. To account for the observations made we perform follower pattern analysis, using following coefficient to quantify followership strength of an author with

respect to another author. Conjecture was that rise in citations of sleeping beauty in vicinity of being cited by higher rank author is due to the attention that it attracts from his follower through him, but results were not as expected as described in section 5.4.

For **future work**, we look to improve following coefficient by incorporating topic information and other possible behaviors in its calculation. We aim to find other factors that effect the evolution of sleeping beauty, and extend the analysis to other popularity patterns mentioned in Chapter 1.

Bibliography

[1] E. Garfield. *Essays of an Information Scientist: 1979-1980*. Essays of an Information Scientist. ISI Press, 1977. ISBN 9780894950124. URL https://books.google.co.in/books?id=0ocnAQAAIAAJ. 5

[2] Eugene Garfield. Delayed recognition in scientific discovery: citation frequency analysis aids the search for case histories. *Current Contents*, 23:3–9, 1989, Reprinted in *Essays of an Information Scientist: Creativity, Delayed Recognition, and other Essays*, 12:154, 1989. 5

[3] Qing Ke, Emilio Ferrara, Filippo Radicchi, and Alessandro Flammini. Defining and identifying sleeping beauties in science. *Proceedings of the National Academy of Sciences*, page 201424329, 2015. 5, 6, 17

[4] Christian Lachance and Vincent Larivière. On the citation lifecycle of papers with delayed recognition. *Journal of Informetrics*, 8(4):863–872, 2014. 5

[5] John Mingers and Loet Leydesdorff. A review of theory and practice in scientometrics. *European Journal of Operational Research*, 2015. 4

[6] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: bringing order to the web. 1999. 10

[7] Hendrik P Van Dalen and K? ne Henkens. Signals in science-on the importance of signaling in gaining attention in science. *Scientometrics*, 64(2):209–233, 2005. 5

[8] Xiao Yu, Quanquan Gu, Mianwei Zhou, and Jiawei Han. Citation prediction in heterogeneous bibliographic networks. In *SDM*, volume 12, pages 1119–1130. SIAM, 2012. 3