Analysis of Citation Patterns in Society of Statistics

Minshi Peng
Advisor: Jiashun Jin(CMU), Pengsheng Ji(U Georgia)

May 15, 2017

1 Introduction

Significance In the past decades, the scientific community has grown substantially: we have way more researchers and annual publications than we ever had before. While great progress has been made in our time, there are also many critical criticisms, like the paper by [?], Why most published research findings are false, and the paper by [?], Opinion: Science in the age of selfies. An interesting question is therefore how to scrutinize the vast volume of scientific research we have today.

Describe the Status Quo In the last years, due to the increasing availability of computational resources, it is possible to build large bibliographic data set and analyze them to an unprecedented level of accuracy. Redner [?] performed an analysis over all papers published in the 110 years long history of journals of the American Physical Society (APS), the citation data cover 353,268 papers and 3,110,839 citations. Another well-studied dataset *CiteSeer* contains over 800,000 research papers in computer science published by over 2 million authors. Some patent-citation data has also been collected and cleaned, like *cit-patents*[?], which contains 3,774,768 papers and 16,518,948 citations.

Identify A Gap Studies on these large bibliographic data sets cast light on the internal structure of the research communities, and provide us with a better understanding of the historical roots and development of those research fields. Though statisticians contributed a lot in studying those bibliographic data, however, till now few work has been done to study the structure of statistics society, that is the purpose of my project. Our data sets covers about 90,000 research papers published in 36 journals in statistics from 1976 to the first half of 2015, consisting of titles, authors and affiliations, abstracts, MSC numbers and keywords, etc. The Phase I of the data, containing papers published in the top 4 journals in statistics in last ten years, has recently become public[?]. The work has drawn unexpected attentions and was considered as a great step forward to provide the community with a first such data set for self-study. The Phase II data is a step further, with the increased breadth and width, opens many potential more interesting questions. One is related to the citation network, which is the central concern of my ADA project.

Fill That Gap With Present Research Citation networks are compact dynamic representations of the relationships between research products. they offer a fertile ground for studying research and collaboration patterns of scientific communities. Basically, we wish to understand why paper get cited and why paper get highly cited. We want to study the citation patterns and how the citations evolve over time, and what features of a paper makes it highly cited. Hopefully, prediction model can be built to estimate the influence of the research result on the current community before it is published, or even be generalized and applied to other fields, like citation pattern prediction. A better understanding of such citation features is useful in many perspectives: it may help administrators or funding agencies to prioritize research areas, and researchers to start a new topic or a publish a new paper.

2 EDA

Our dataset covers all papers from 36 journals in 1976-2014 and half of 2015. The total number of papers is 80989. After removing the papers with missing information(those are letters to editors, book reviews and paper discussion) and the duplicated records, there are in total 71918 papers, which results in 390785 citations among these papers. The list of 36 journals covered in this data set is shown in table 1. These journals contains the major influential Statistical journals and several closely related important journals in other fields.

Poincare	AoAS	AoP	AoS	AISM	AuNZJS	BayN	Bern
Bmtrcs	Bmtrk	Biost	CaJS	ComSTM	CSDA	EJS	Extrm
ISRev	JCGS	JMLR	JASA	JRSS-B	JoAS	JClass	JMVA
JRSS-A	JRSS-C	JSPI	JoTS	JoNpS	PTRF	StSci	ScanJS
Sinica	StComp	SPLet	StMed				

Table 1: 36 journals covered in our data set(in shorhand)

The data set also provides detailed information of each paper, including:

- Identification number: MR/DOI/WOS number, so that each paper can be identified uniquely.
- Authors: The authors and coauthors for each paper.
- Publication information: Journal, Year, Total number of pages, Volume and issue, Url.
- Title, Abstract: The full title and abstract for each paper.
- Citation list: The reference list of each paper is extracted in form of WOS number.

Figure 1 left shows the histogram of papers published in each year from 1976 to 2015. The abrupt decrease of the last bar is due to the half portion of data we have for 2015, thus the height of the bar is expected to be twice the amount, which is consistent with the trend of previous years. We can see the number kept increasing over the years, and increased most rapidly during 2000-2008. This is probably due to the growing interests in ”Large Data”. The right figure shows the average number of citations received till 2015 of papers published in each year, i.e. the y axis represents how many citations received by the papers published in this year till 2015, divided by the total number of papers published in this year. The general trend is that the number keeps going down with some fluctuations. This makes sense since averagely older papers have more time to be discovered, explored and cited.

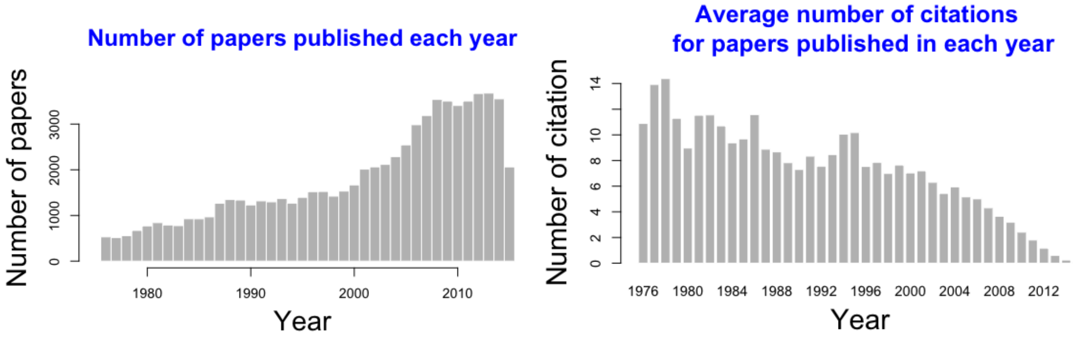

Figure 1: Left: histogram of papers published in each year from 1976 to 2015. Right: The average number of citations for papers published in each year till 2015.

Figure 2 shows the citation distribution for two groups of papers. The left figure shows the number of citations for papers published during 1976-1980 and the left figure shows citations for papers published during 1981-1985. Since in each figure the papers are published roughly at the same time, their citations are comparable. We can conclude from the histogram that the citations per paper follow roughly a power law distribution, that the number of papers drops exponentially as the citation counts increase. According to the two histograms most papers have less than 10 citations within the 36 journals till 2015 while few papers have hundreds of citations.

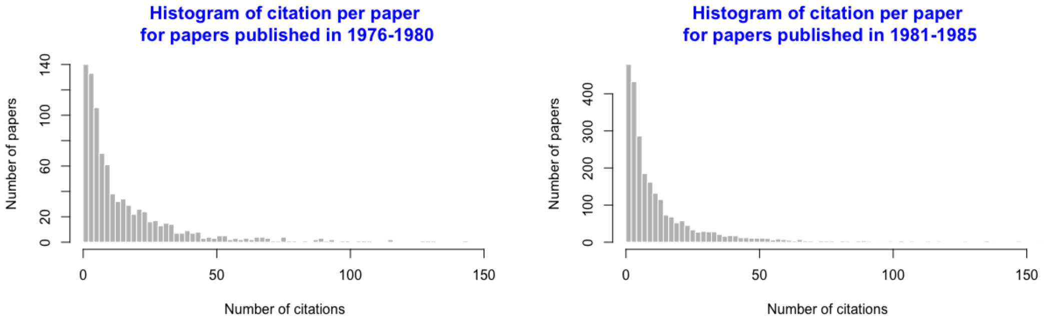

Figure 2: Histograms of number of citations each paper received. The left histogram is restricted to paper published during 1976-1980 and the right figure is restricted to papers published during 1981-1985.

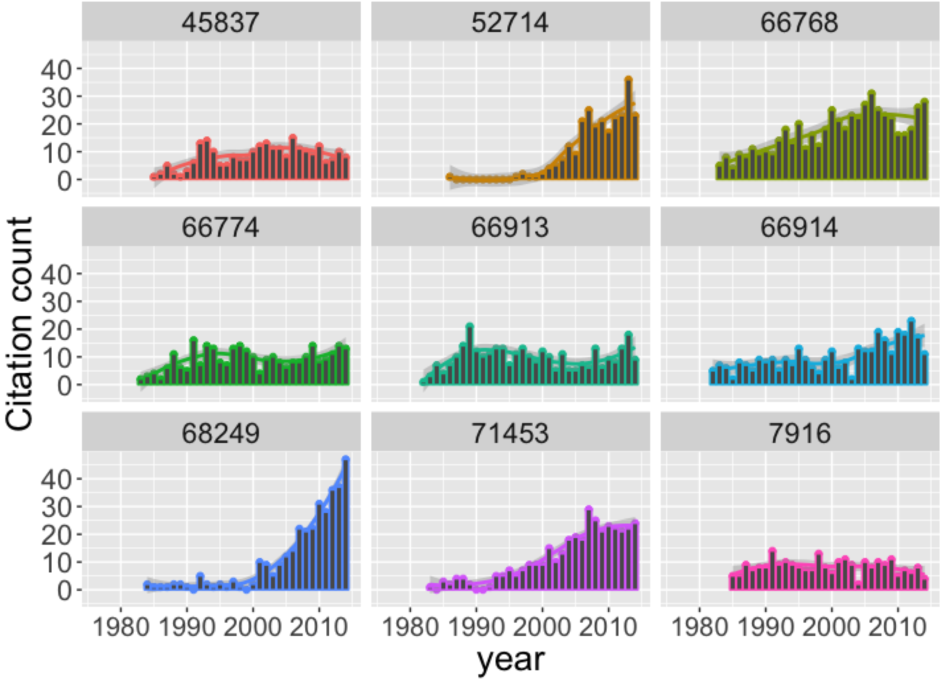

Figure 3: Citation patterns for 9 highly-cited papers published during 1981-1986

Next we give an example of how citation pattern can be varied for different papers. Figure 3 shows the citation patterns of 9 highly-cited papers published during 1981-1986, where the x axis shows the year and y axis is the citation for each year. Though these papers are all comparatively highly-cited and published at the same time, the citations grew distinctly over the years. Some papers, for example No.52741, acted as a sleeping beauty. It kept silent for almost 15 year and then suddenly became popular. Some other papers, for example No.66768, have a steadily increasing citations, while some papers had most of it's citations right after its publication, and then the popularity died out over the year.

3 Analysis

The number of citations appeared each year for each paper is a complex mechanic that evolve many factors, considering all the variables at the same time seems to be an impossible mission. To begin with, we consider some potential factors separately to see whether they have an interrelation with the number of citations.

3.1 Co-authorship

We can observe that there are many papers that have more than one authors. Indeed, in academia the collaboration is very common since each scholar has his/her strength and weakness. Therefore the number of authors of the paper serves as an potential factor that determine the citations of the paper. So we are interested in what proportion of the papers having one, two, three or four and more authors, and are there any difference between highly-cited papers and averagely-cited papers. Since the co-authorship might have been changing over time, we cut the 40-years period from 1976 to 2015 as 10 equal time segments. So papers published in each time segments can be view as appearing at the same time, and therefore their citation counts are comparable. First we calculate the proportion of different co-authorship for papers published in each time period. The result is shown in left panel of figure 4. We can observe that over the years, the proportion of single-authored papers kept decreasing, while the proportion of papers having two, three or four and more authors were all increasing, and among them the proportion of three-author papers increased comparatively faster.

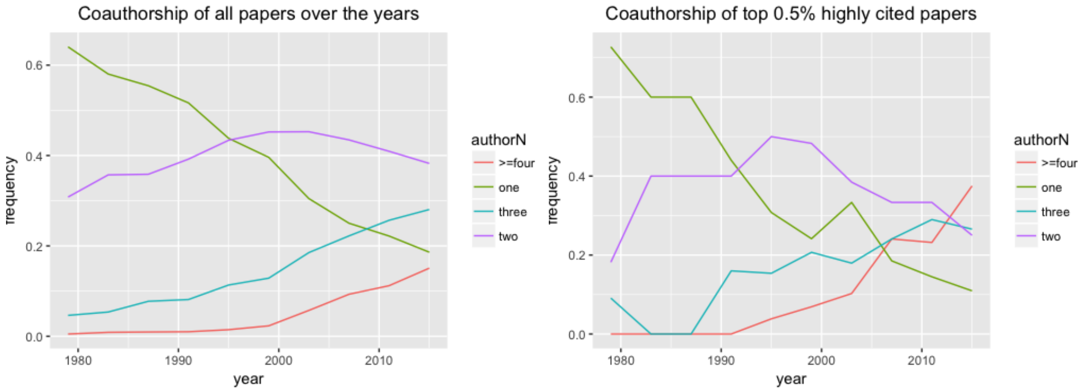

Figure 4: Left: the proportion of general papers having one, two, three or four and more authors, for each one of ten period of time from 1976 to 2015. Right: the proportion of each number-of-author for the top 0.5% highly-cited papers for each time period.

1976-1979	1980-1983	1984-1987	1988-1991	1992-1995
11	15	20	25	26
1996-1999	2000-2003	2004-2007	2008-2011	2012-2015
29	39	54	69	64

Table 2: The number of papers having top 0.5% number of citations among papers published in same four-year period.

Then we select the top 0.5% highly-cited papers during each time period, and summarize the distribution of the number of authors as proportion.Table 2 shows the number of papers selected for each four-year period. Figure 4 shows the proportion of those papers having different number of authors. The general trend over the years follows the similar pattern, that the number of single-authored papers decreased sharply while the multiple-authored papers were gaining increasing popularity. However the difference is also evident between two patterns. The most prominent difference is the proportion of papers having four or more authors, which eventually became dominated among the top-cited papers, while it remained modest among the average papers. It seems the highly-cited papers prefer more authors, and this makes sense because creativity is often sparked by the collision of different ideas. We can probably conjecture that the proportion of multiple-authored papers would continue to increase, especially among the highly-cited papers.

Hypothesis Apart from the number of authors, we are also interested in the composition of the authors if the paper have more than one authors, and whether this is closely related to the citation behaviors. We observe that there are different kind of co-authorship, for example, PhD student and his/her advisor, co-workers in the same university or scholars from different institutes. Since this requires the information of the authors which we don't have currently, this remains to be an hypothesis.

3.2 Keywords in title

Next we consider whether the topics of the papers are associated with the number of citations, and how did the relation change over time. We measure the association by taking correlation of the existence of this topic in the title of papers and the number of citations the papers had in each period. We consider the citations appeared in a sequence of two-year period from 1980 to 2013, so that we could measure how the correlations evolved over the years.

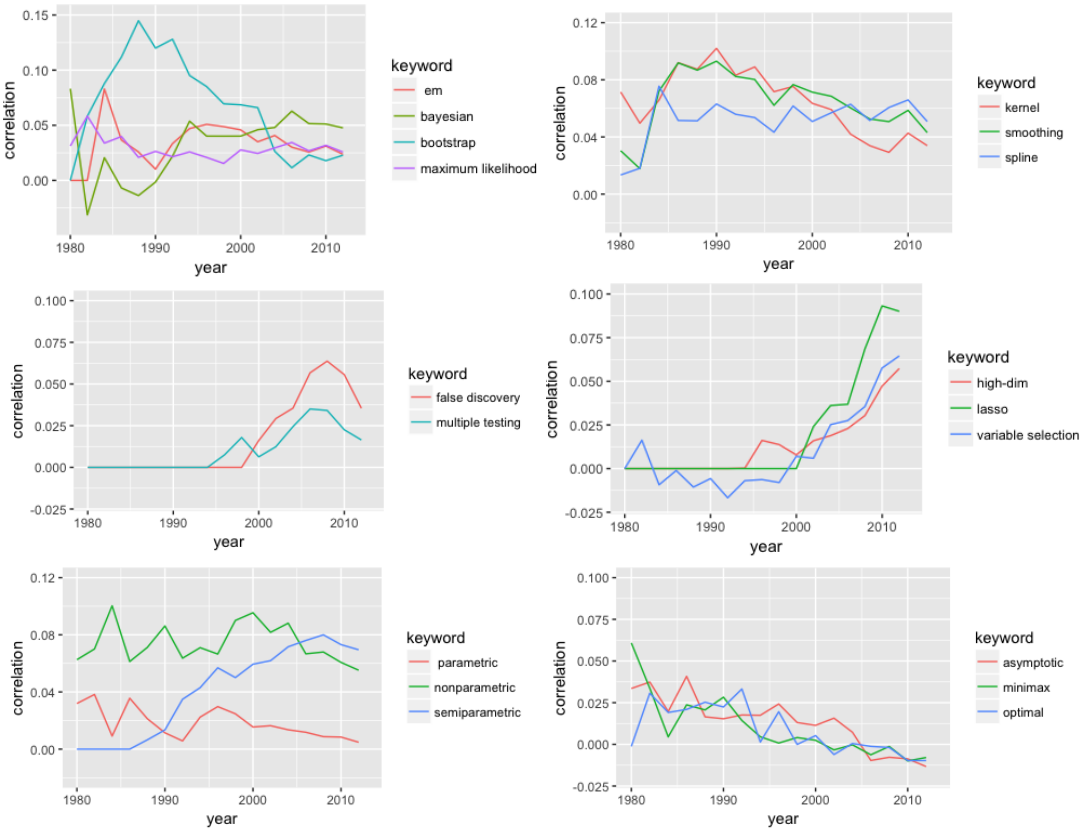

Figure 5: The correlation of existence of a keyword in the title with the number of citations the papers received in each two-year period, ie, how the topics had an impact on the number of citations of the papers over the years

Figure 5 shows some representative topics that had non-trivial correlations with the citations. The keywords in the first four figures covers some hot topics in Statistics. We can clearly find some trend in these patterns, and this, to some extent, concords with the popularity of the these areas in society of Statistics over the years. For example, the pattern of keyword 'bootstrap' in the first figure shows that in 1990 papers about 'bootstrap' were averagely receiving higher citations, which agreed with the mainstream research interest at that time. The first figure also shows that the 'bayesian'-related topic once had an negative effect on the citations, however, later the impact had been reversed to a nontrivial positive impact till now. The second figure shows that the 'smoothing'-type of topic was most popular and highly cited in 1990s and the positive impact continued to recent years, though slightly weakened. The third figure shows the popularity pattern of multiple testing problems, which were highly-cited around 2006. The forth figure shows some of the most popular topics for the last ten years, like 'high dimensional', 'lasso' and 'variable selection'. The fifth figure shows the impact on citation by which class of the papers belongs to, parametric, nonparametric or semi-parametric. Averagely nonparametric papers were associated with higher citations than the parametric papers, and the semi-parametric papers seems to be a rising star, which was becoming popular. We can possibly guess this topic would becomes more popular in the near future. The last figure shows that some theoretical topics resulted in comparatively lower citations. This seems to be counterintuitive at first, since most of these papers were written by renowned statisticians. But it also makes sense, since the theoretic papers always serve as ’topic killer’, because it marks that this kind of problem has been thoroughly done, with little opening for future work. By analyze the correlation of citations and the keyword, we can gain some insight in how the research interests in statistics society changed over the years.

3.3 Journal

Our data set covers papers in 36 journals in Statistics and some closely related fields. It is conceivable that papers published in more famous journals would probably receive more citations. However, though there exists various evaluation of different journals, there is no universally acknowledged ranking among them. In addition, the popularity of the journals keeps changing over the years. Therefore we want to analyze how the citation behavior of the papers associated with the journals where they were published, and how it changed over time. We consider 9 of currently most popular journals: Annals of Applied Statistics(AoAS), Annals of Statistics(AOS), Biometrics(Bmtrc), Biometrika(Bmtrk), Computational Statistics & Data Analysis(CSDA), Journal of the American Statistical Association(JASA), Journal of Machine Learning Research(JMLR), Journal of the Royal Statistical Society Series B-Statistical Methodology(JRSS-B), Statistics in Medicine(StMed). Figure 6 left shows the total number of papers published in each journal. Some of the journals were established later, like JMLR, so the total number of papers are less than the others. Figure 6 right shows the average number of citations of papers published in each journal. We observe that JRSS-B has the highest average citations, followed by JASA, while JoAS, CSDA and StMed has comparatively lower average citations.

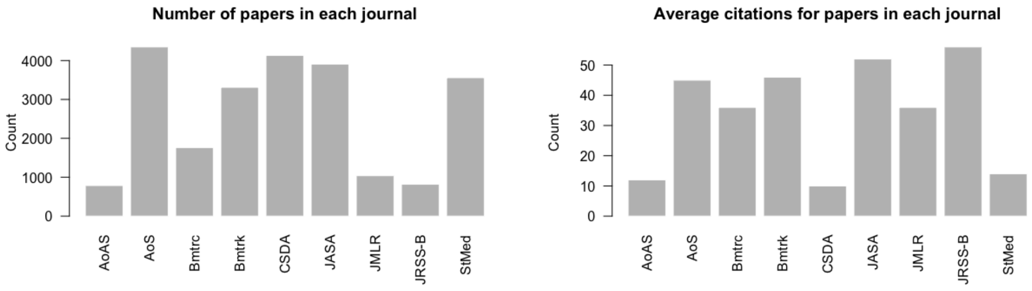

Figure 6: Left: the total number of papers published in each journal in our data set. Right: the average number of citations of papers in each journal.

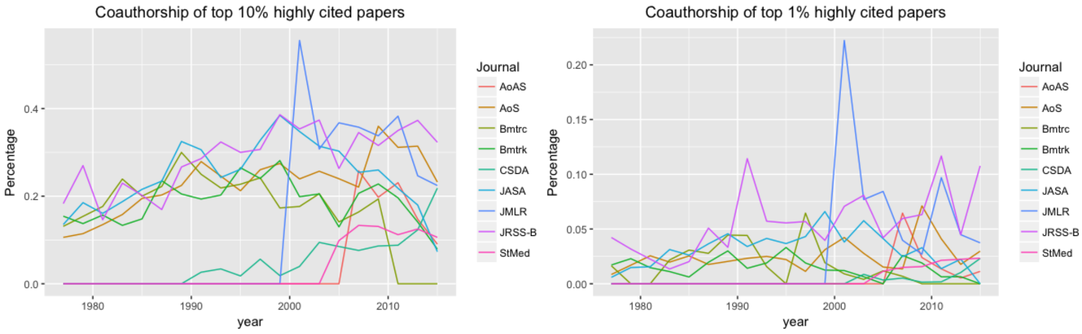

Figure 7: For papers published in each period of time, the percentage of papers in each journal that ranked among the top in term of citations. In the left figure we consider top 10% while in right figure we consider the top 1%.

The average citation can tell some story, however, the chances are that it could be the few ”star-papers” that contributed the most citations. Further, we are also interested in how the behavior changed over the years. So again we cut the 40-years time span to 10 segments, and consider the papers published in each time period separately. For papers published in each period and each journal, we calculate the percentage that has been highly-cited, i.e. the number of citations ranks top 1% or 10%. The results are shown in Figure 7. The patterns after 2010 are erratic, since those papers were recently published, so there are too little citations to distinguish between the papers having high citations and low citations. We can observe that, when considering the top 10% highly-cited papers, the five journals including AoS, Bmtrk, Bmtrc, JASA and JRSS, all maintains high percentage of well-cited papers in early years. Then after 2000, the pattern began to diverge, where the percentage of high-cited papers of AoS, Bmtrk and Bmtrc dropped, while there were some rising star like JMLR, CSDA and AoAS, that were having an increasing portion of highly-cited papers. For the pattern shown in the right panel, JRSS-B stands out for its large percentage of the top-cited papers. Also the performance of JMLR is noticeable since it remains a high-proportion of top-cited papers after 2000. We can therefore conclude that the citation of papers is related to the journal it is published, and the relationship differs in items of time and level of citation.

3.4 Citation Patterns

As is shown in Figure 3, the citation pattern varied from papers to papers. However, it seems the patterns for highly-cited papers also share some common trend. We observe that for older highly-cited papers, most citation patterns follow figure 8, where the citation per year remained at a low level till the beginning of 21st century, and from which the number started to increase rapidly. The main difference between those citation patterns lies in the weight on the first and the second part. Some of the patterns are light-tailed and others are heavy-tailed.

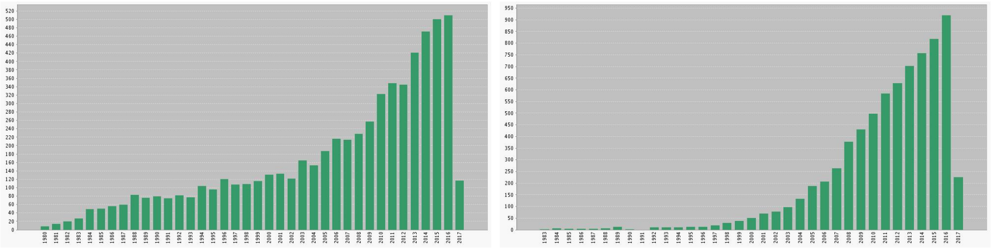

Figure 8: The Typical citation patterns for the highly cited papers in early times (around 1980s).

More recent highly-cited papers have more varied citation patterns. Two typical patterns are shown in figure ??. The pattern in the left panel is the most common one, in which the rate increased (over or nearly) linearly till around 2008-2010, then the citations per year stopped increasing, instead, for some paper it began to drop. The other pattern shown in right panel represents the papers having the number of citations increased soon after its publication, then the popularity stayed at a high level till about 2010, and began to decrease slowly afterwards.

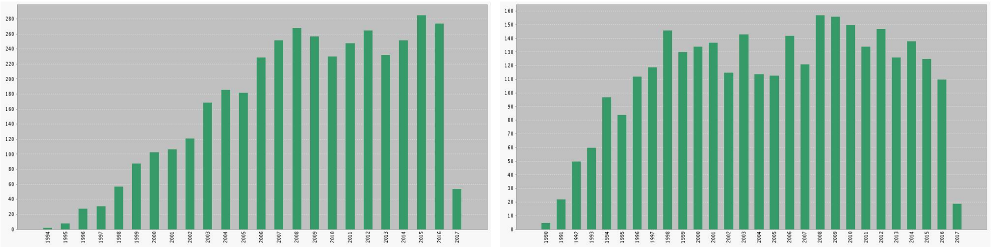

Figure 9: The Typical citation patterns for middle-aged highly cited papers (around 1990s)

Hypothesis: 2000 and 2010 might be two historical change points. Something happened in 2000 that witnessed the beginning of prosperity of those papers, while some other reasons arose around 2010 that marked the saturation of the citation activities, thus the citation per year of the papers stopped increasing. There might be multiple contributors, for instance, there are more institutions that publish papers, or the amount published in each journal increased that resulted in the rise in the total number of papers each year; or the citation preference has changed, so the authors tend to cite more(or less) papers to increase their citations. The more detailed analysis for these hypothesis remains to be our future work.

4 Future plan

- Continue the analysis on the keywords. Find the keywords that is becoming popular recently.
- Analyze other factors, including the influence of the authors(citation indices of the author), the MSC classification subjects, etc.
- Construct the regression model to consider all the variables together.

References

[Geman and Geman, 2016] Geman, D. and Geman, S. (2016). Opinion: Science in the age of selfies. *Proceedings of the National Academy of Sciences*, 113(34):9384–9387.

[Ioannidis, 2005] Ioannidis, J. P. (2005). Why most published research findings are false. *PLos med*, 2(8):e124.

[Ji et al., 2017] Ji, P., Jin, J., et al. (2017). Coauthorship and citation networks for statisticians. *The Annals of Applied Statistics*, 10(4):1779–1812.

[Leskovec et al., 2005] Leskovec, J., Kleinberg, J., and Faloutsos, C. (2005). Graphs over time: densification laws, shrinking diameters and possible explanations. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 177–187. ACM.

[Redner, 2005] Redner, S. (2005). Citation statistics from 110 years of physical review. *Physics today*, 58(6):49–54.