Diffusion Pattern of False Discovery Rate

Ying Chen Xiaotian Wang Xiyu Yang Shuqi Yu

Abstract

In this project, we study on one specific technique, false discovery rate (FDR), analyzing the citation network to explain how this technique is broadly used and affects the academic world. With easily-accessible, detailed data from Semantic Scholar, we build the citation network for FDR in the past 26 years. Analysis on the network enables us to understand the diffusion pattern of this technique. We find the overall diffusion pattern and investigate the popular fields of study using FDR. In this study, we conclude that, first, no overall time lag exists for use of FDR, however the lags exist in application fields outside of statistics. Second, the influence keep expanding in the past 26 years, but the increasing rate slows down in the past 5 years. Third, the diffusion patterns are different across the fields, and we find evidence of 'bridge' paper in the fields outside of statistics.

1 Motivation and Goal

Researchers are interested in tracking how the citation pattern of a certain work changes over time or among fields of study. They want to know if their proposed techniques are widely accepted and influence outside their own academic field, or even just out of curiosity. There are works on citation networks considering hot fields and combining with coauthorship network to investigate the citation pattern in general Radicchi et al. (2012) and Ji et al. (2016). However, to our knowledge, very few work (if any) focuses on a specific statistical technique, tracking its citation in different fields of study along time.

In this project, we study one specific statistical technique and its citation diffusion pattern. We would like to understand how citation changes over time and among different fields. Specifically, first, we observe the time lag after the first work was published, comparing across the fields. Second, we explore the shape of the diffusion patterns in each field. Third, as we focus on a statistical technique, we also observe the difference of the diffusion patterns across the fields, based on their relevance to statistics. We try to shed some light on the study of citation diffusion pattern.

2 Data collection

We use the Semantic Scholar's records, which gives information on research papers. The corpus Ammar et al. (2018) is composed of rich abstracts, bibliographic references, and structured full texts. The full text comes with automatically detected inline citations, figures, and tables. Also, each citation is linked to its corresponding paper. In Semantic Scholar's dataset, papers from hundreds of academic publishers and digital archives are aggregated into a single source. With the well-developed tool, we download and configure the whole dataset with 220 million papers and over 100Gb meta data. They are well prepared, machine-readable academic texts.

The huge database gives us a chance to explore citation diffusion pattern in depth. In this study, we focus on false discovery rate(FDR), trying to find the relationship among the papers that use this technique.

False discovery rate is first introduced by Benjamini and Hochberg (1995). FDR conceptualizes the rate of type I errors in null hypothesis testing with multiple comparisons and FDR-controlling procedures have greater power, at the cost of increased numbers of Type I errors. It gains broad acceptance in many scientific fields including life sciences, genetics, biochemistry, oncology and plant sciences.

2.1 Dataset I

From the Semantic Scholar's records, we pick all research papers that mention FDR in their abstracts. We scan through the raw dataset and include the paper containing the key phase 'false discovery rate' (not case sensitive) in the abstracts. We bring in the variables paper ID, title, abstract, year published, citation relationship, journal name, and field of study of each paper. For each paper, citation relationship includes two parts, inCitation and outCitation. InCitation is a list of paper IDs which cited this paper, while outCitaion is a list of paper IDs which this paper cited.

2.2 Dataset II

We rebuild the dataset in a different way, containing all the papers that directly cited the original work Benjamini and Hochberg (1995) in the raw data. We use the same set of variables as in dataset I for each paper.

3 Analysis

3.1 Analysis on dataset I

We get 8,787 papers after filtering. The size is comparable small with dataset II, however we get more accurate knowledge on the paper using FDR on this dataset and the size is enough for us to observe some interesting features of citation pattern.

3.1.1 Overall trend by year

We use citation ratio to determine popularity of the technique, i.e. the number of paper in dataset I over the total number of publication in each year. The Benjamini and Hochberg paper is released in 1995. We find a clear increasing trend along time. The ratio start to increase rapidly around 2002. It fluctuates at a high level from 2012 to now. As we don't have enough observation, it's hard to conclude if FDR is less used in the past few years.

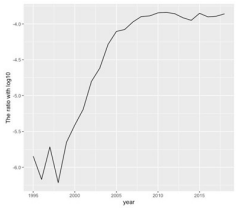

Figure 1: The overall citation trend of FDR under the log scale.

3.1.2 Citation adjacent matrix

To explore how FDR is used in different fields, we construct the adjacent matrix of citation network for both in and out citation. The rows and columns of the matrix represent the papers in dataset I. In-citation matrix A is defined as

$$A_{ij} = \mathbf{1}\,(\text{Paper i is cited by paper j})$$

and out-citation matrix B is

$$B_{ij} = \mathbf{1}\,(\text{Paper i cites paper j})$$

3.1.3 Clustering by VSP

We use VSP to analyze the above matrices. VSP is a spectral method that combines principal components analysis (PCA) with the varimax rotation. This algorithm is introduced by Rohe and Zeng (2020). Under mild assumptions, the VSP estimator is consistent for degree-corrected Stochastic Block models. Roughly, the algorithm contains three steps: centering the matrix; applying singular value decomposition(SVD) to get the top k left and right singular vectors; rotating these singular vectors to achieve the maximize Varimax. The optimal centering step speeds up the algorithm when the matrix is sparse.

To determine the rank in VSP, we first assign a pre-picked number and show the scree plots for the matrices. After trail-and-error, we use rank 6 for A (in citation) and 5 for B (out citation) respectively, see Figure 2.

Both plots have a gap between the third and the forth eigenvalue. So, there are at least 3 reasonable clusters here. We can pick k = 3, and plot their top three principal components,

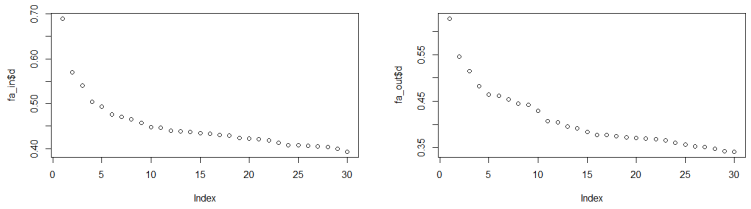

Figure 2: Scree plots with rank 30 for matrix A and B.

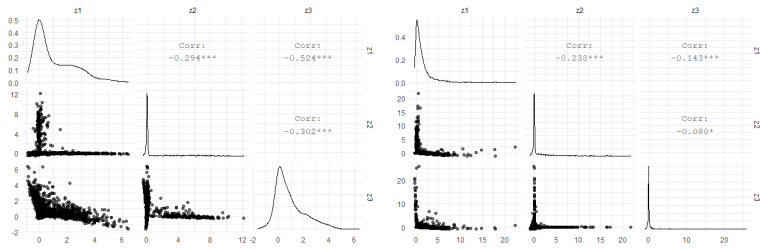

Figure 3: Scatter plots of the three leading principal components for in and out citation. The figure on left corresponds to the inCitaion adjacent matrix A, the figure on right corresponds to the outCitaion adjacent matrix B.

see Figure 3. In the figure for in-citation, each scatter plot shows a clear L-shape. However in the figure for out-citation, the scatter plot on the bottom left corner doesn't have a prefect L-shape, which indicates the rank greater than 3. Indeed, it's unreliable to guess the rank k by simply observing the gap on the scree plot.

3.1.4 Contextualization by bag-of-words

We contextualize the clusters by analyzing the paper abstracts using bag-of-words(bff), and we don't mind the eigengap which is inspired by Wang and Rohe. Here, we construct two new matrices for in and out citation network. We use the matrix as an external information to illustrate the features of clusters in paper-inCitations network and paper-outCitations network.

Use the adjacent matrix $\tilde{A}$ and $\tilde{B}$ to denote paper-abstract network for in and out citation. To construct this matrix, we first remove all the numbers from the abstracts in the data set. Next, we convert abstracts into tokens (words). We then remove the most common words from the list. In the matrix, we have each paper as a row and a single word as a column.

For the inCitation network, we find 3 meaningful clusters, see Figure 13 in Appendix for details. Based on the key words, we summarize them as *statistics, proteomics*, and

genetics.

Similarly, we find 7 meaningful clusters for out citation network as *hypothesis testing, proteomics, gene expression, genetic variations in human, genetic variations in plant, regression*, and *DNA methylation*, see Figure 14 in Appendix for details.

3.1.5 Update rank for outCitation network

Inspired by the bag-of-words results above, we redo the VSP on the outCitation adjacent matrix with rank 7, see Figure 4.

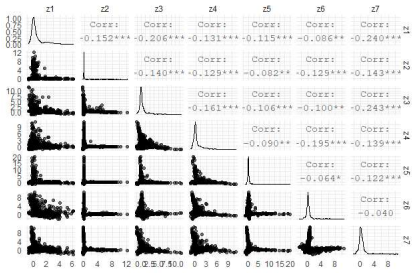

Figure 4: The scatter plots for the seven leading principal components corresponds to the outCitaion adjacent matrix B.

3.1.6 Citation patterns across clusters

We study on the 5,182 papers are included in the above 7 meaningful clusters from the outCitation network. For each cluster, we plot the trend over time similar to the overall trend we have in Section 3.1.1 in Figure 5.

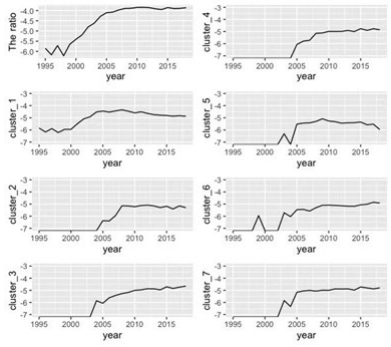

Figure 5: The trend for each cluster(on OutCitation Network).

Firstly, the ratio remains at a high level in all the plots which indicate the influence of FDR is still expanding. And we observe different patterns across the clusteers. Time lag

appears in cluster 2 (*proteomics*), 3 (*gene expression*), 4 (*genetic variations in human*) and 7 (*DNA methylation*). In cluster 1(*hypothesis testing*), FDR is popular shortly after the release and thus no time lag.

3.1.7 Visualization of clusters

With the clusters identified, we visualize them to see their behaviors and connections across the clusters. In each cluster, we rank papers by loading. The higher the loading is, the closer the paper is related to the cluster. To make the plot clearer, we pick top 10% and top 50 papers. We see more citations within a cluster than between two clusters, see Figure 6.

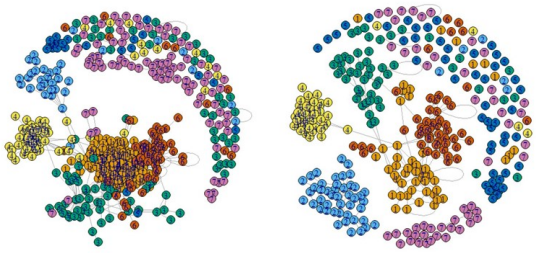

Figure 6: The visualization on outCitation network with 7 clusters. The left figure plots the top 10% papers, the right one plots the top 50 papers in each cluster. Each dot denotes a paper, while the line denotes the citation relationship.

We further simplify the graph (Figure 6) by presenting each cluster as one single dot to get a clear view, see Figure 7. It's clear in the updated graph that cluster 1 (*hypothesis testing*) contains the majority of papers. Cluster 6 (*regression*), which is closer to hypothesis testing, has a larger citation number compared to other clusters. And since we use out citation for the graph, the most citation are from hypothesis testing to others.

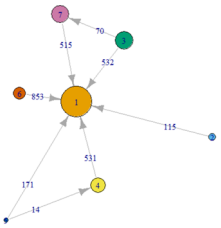

Figure 7: The main structure of outCitation network with 7 clusters. The size of the dot indicates number of papers in the cluster. Also, we only keep the edges in which number of citations of the dot counts for more than 5% total citations..

3.1.8 Bridge paper

In the visualisation result, we see more citation within the cluster than across clusters. This bring us to the question whether there exists some 'bridge' paper, which means they don't cite the origin work directly when using the technique. We define bridge paper as influential paper other than the origin work that other papers cites a lot in a certain field.

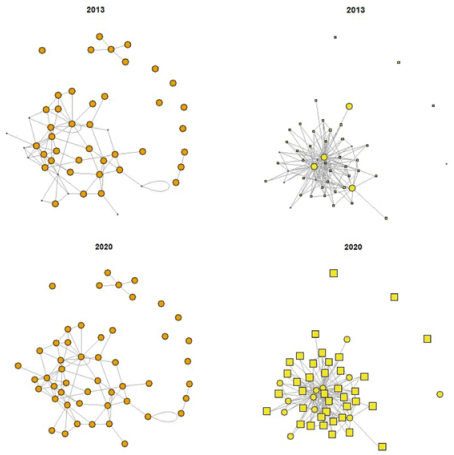

Figure 8: The two figures on the left represent the network for cluster 1 (Hypothesis testing). The two figures on the right represent the network for cluster 4(Genetic variations in human). The round dot indicates the paper cited the original paper while the squared dot indicates the opposite. This Figure is based on Out-Citation network with 7 clusters.

Comparing hypothesis testing with human genetics, we find that the clusters closely related to statistics are more likely to cite the origin work directly while those clusters outside of statistics are more likely to have a bridge paper in their own field.

3.2 Analysis on dataset II

Total number of papers is 41,067 in dataset II. In this larger dataset, we expect to observe more clusters than previous one. We apply the similar analysis steps in section 3.1.

3.2.1 Resulting Clusters

Based on the out-citation network, we find 5 meaningful clusters include 39,985 papers, see Figure 9 for their citation ratio. We see some upward trends in the cluster *gene expression*, *neurosciences* and *microbiology*.

In the analysis of in-citation network, we have 12,328 paper in 7 meaningful clusters, see Figure 10. In some clusters like *feature engineering*, *gene expression*, and *radiology*,

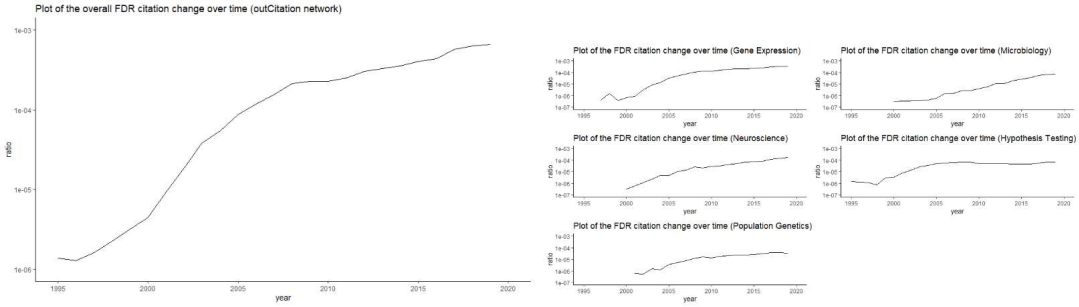

Figure 9: The trend base on out-citation network under the log scale.

the trend doesn't appear for the whole period, implying the different popularity of FDR across fields. This give us some clue of the bridge paper. But based on the main result, it's vague. Another interesting finding in in-citation network is that a cluster for *wine making* pops out. FDR, as a cutting-edge tool, benefits researches in different disciplines.

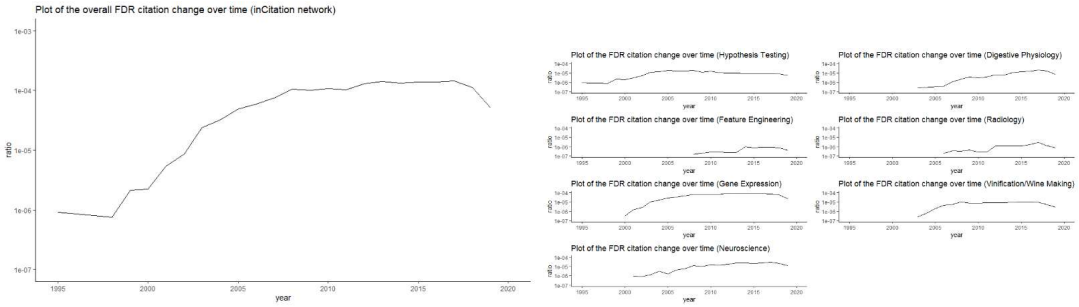

Figure 10: The trend base on in-citation network under the log scale.

4 Conclusion

In the study, we identify the citation diffusion trend of FDR and analyze the pattern in different fields. From the analysis above, we find almost no time lag for FDR overall. Citation started shortly after the technique came out, which means FDR becomes popular fast. However, time lags widely appear in application fields outside of statistics such as genetics, engineering and wine making.

Overall, the FDR remains hot and it keeps in a high level under the log scale, however the influence increasing rate slows down in the past 5 years. The diffusion patterns are different across the fields, but mostly remain increasing with different slops.

And we observe some evidence of bridge papers in the fields outside of statistics, but we don't have detail information about them in our dataset.

A Trial on a different technique

In another trail, we use LASSO as key word, and journal names as variable of interest. We would like to see if certain technique appear on some journals more often than others. Similarly, we scan all abstract and pull out papers. There are 9,944 research papers in 4,281 different journals.

Journals or archives with most papers that mentioned LASSO are ArXiv (290), asXiv: Methodology (160), arXiv: Statistics Theory (154), PLoS ONE (126) and bioRxiv (82). Among the 4,281 journals, there are 262 journals/archives with more than 5 papers that mention LASSO.

	journal	Frequency
	<chr>	<int>
1	ArXiv	290
2	arXiv: Methodology	160
3	arXiv: Statistics Theory	153
4	PLoS ONE	126
5	bioRxiv	82
6	Scientific Reports	64
7	International journal of systematic and evolutionary microbiology	62
8	Comput. Stat. Data Anal.	60
9	arXiv: Machine Learning	57

Figure 11: Paper distribution in journals/archives.

From the paper-word pair, we create a matrix with paper ID as row and each word as column. Then, we use VSP to see if there are any factors standing out. Based on the output, instead of gathering by journal names, papers are clustered by languages. However, there are some paper with more than one abstract in different languages. It's hard to get rid of the influence from different languages. In our result, we see factors from both technical topics and languages. Cluster 1 and 3 shows technical topics while 2 and 4 are influenced by languages.

[,1]	[,2]	[,3]	[,4]
"regression"	"<U+6F0F>"	"cancer"	"<U+55A2>"
"selection"	"<U+8119>"	"patients"	"<U+9225>"
"sparse"	"de"	"of"	"<U+707B>"
"lasso"	"des"	"signature"	"<U+63B3>"
"dimensional"	"s"	"prognostic"	"factors"
"variable"	"la"	"carcinoma"	"influencing"
"methods"	"pour"	"cell"	"china"
"data"	"et"	"radiomics"	"<U+7286>"
"models"	"es"	"predicting"	"<U+59D1>"
"linear"	"en"	"predict"	"<U+9E7F>"

Figure 12: Key words in clusters.

Analysis based on journal names doesn't give a decent result. So in the main analysis, we use in and out citation relationships.

B Extra Figures

In this section, we attach more figures.

V1	V2	V3
hypotheses	peptide	conditional
procedures	identifications	gwas
null	peptides	loci
procedure	spectra	shared
controlling	proteomics	pleiotropic
testing	search	summary
proportion	decoy	wide
error	mass	overlap
problem	database	pleiotropy
statistics	matches	enrichment
benjamini	spectrum	genetic
hochberg	tandem	epidemiological
true	spectrometry	genome
proposed	ms	schizophrenia
power	identification	polygenic
rejections	protein	phenotypes
rejected	proteins	statistics
dependence	shotgun	association
simulation	engines	conjunctional
paper	proteomic	snps

Figure 13: The bag-of-words results for inCitaion network with $k = 3$ on dataset I. Each column contains the top twenty representative words in the cluster.

V1	V2	V3	V4	V5	V6	V7
hypotheses	peptide	seq	snps	linkage	lasso	methylation
null	proteomics	sequencing	gwas	disequilibrium	selection	cpg
procedures	peptides	rna	genetic	breeding	knockoff	cpgs
testing	identifications	expression	variants	nucleotide	sparse	dna
procedure	spectra	expressed	association	snp	variables	epigenetic
error	spectrometry	differentially	loci	snps	procedure	methylated
controlling	mass	genes	nucleotide	traits	dimensional	epigenome
multiple	search	gene	wide	marker	knockoffs	beadchip
microarray	database	differential	polymorphisms	association	variable	blood
simulation	ms	reads	genome	population	regression	wide
proposed	tandem	transcriptome	rs	populus	asymptotically	infinium
hypothesis	proteins	read	snp	single	penalized	humanmethylation
proportion	decoy	biological	single	ld	finite	sites
benjamini	protein	transcriptional	traits	trait	gaussian	genes
tests	spectrum	transcripts	risk	phenotypic	power	cg
statistics	identification	edger	phenotypes	polymorphisms	paper	gene
hochberg	matches	regulation	pleiotropy	wood	propose	differentially
power	proteome	regulated	associations	tomentosa	inference	microarray
rejections	engines	pathways	trait	assisted	linear	expression
distribution	spectral	deseq	schizophrenia	plant	prove	cohort
paper	shotgun	throughput	polygenic	markers	control	association
true	mascot	replicates	susceptibility	variation	theoretical	dnam
familywise	proteomic	enriched	shared	genotyped	applications	illumina
probability	searching	degs	conditional	associations	penalty	cord
values	sequest	transcription	disease	mapping	asymptotic	maternal

Figure 14: The bag-of-words results for outCitation network with $k = 7$ on dataset I. Each column contains the top twenty representative words in the cluster.

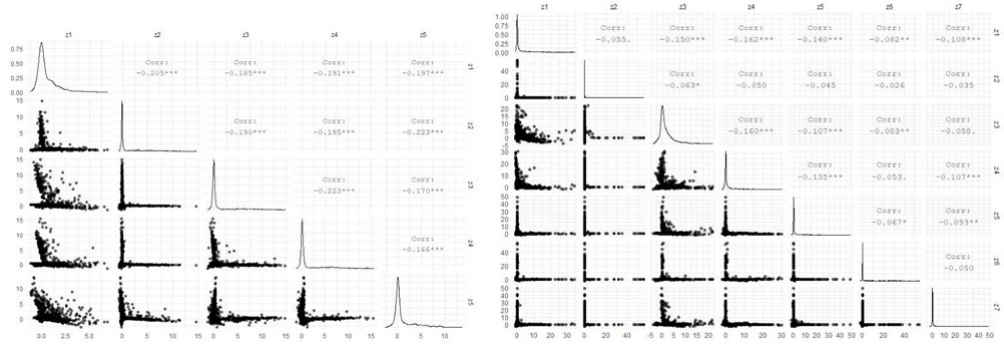

Figure 15: Scatter plot of VSP results base on outCitation network(left) and inCitation network(right) on dataset II.

	V1	V2	V3	V4	V5
1	expression	brain	populations	microbial	false
2	genes	cognitive	population	microbiota	testing
3	genome	imaging	genetic	bacterial	discovery
4	gene	cortex	species	microbiome	procedures
5	cell	cortical	microsatellite	communities	procedure
6	expressed	connectivity	loci	gut	error
7	protein	fmri	dispersal	rrna	statistical
8	transcription	magnetic	diversity	composition	fdr
9	transcriptional	resonance	structure	community	rate
10	transcriptome	neural	evolutionary	bacteria	paper
11	rna	participants	flow	taxa	simulation
12	cells	resting	variation	diversity	null
13	regulatory	frontal	divergence	abundance	proposed
14	differentially	mri	geographic	taxonomic	hypotheses
15	seq	task	selection	sequencing	tests
16	molecular	neuroimaging	habitat	microbes	statistics
17	proteins	eeg	microsatellites	host	method
18	regulation	temporal	conservation	fecal	methods
19	pathways	visual	isolation	genera	multiple
20	identified	subjects	marine	metagenomic	values
21	genomic	parietal	sea	firmicutes	power
22	regulated	prefrontal	traits	bacteroidetes	propose
23	biological	healthy	ecological	otus	estimation
24	cancer	diffusion	history	soil	true
25	cellular	left	north	amplicon	test

	V1	V2	V3	V4	V5	V6	V7
1	false	extraction	gene	brain	microbiota	features	fruit
2	discovery	unsupervised	genes	imaging	gut	radiomics	ripening
3	testing	feature	expression	neuroimaging	microbial	imaging	vitis
4	hypotheses	principal	genome	tensor	microbiome	radiomic	berries
5	fdr	component	http	alzheimer's	composition	tumor	grape
6	rate	fe	microarray	mri	bacterial	images	sauvignon
7	procedure	proposed	biological	connectivity	rrna	image	tomato
8	procedures	recently	wide	cognitive	fecal	cancer	cabernet
9	null	pca	seq	subjects	intestinal	tumors	vinifera
10	error	mirnas	genomic	matter	abundance	prognostic	ethylene
11	benjamini	applied	differential	scans	metagenomic	ct	plant
12	hochberg	drug	sets	cortical	firmicutes	texture	biosynthesis
13	multiple	difficult	rna	regions	host	predictive	fruits
14	controlling	microrna	differentially	diffusion	sequencing	lung	berry
15	statistics	successfully	expressed	magnetic	bacteria	patients	lycopersicum
16	proposed	silico	methylation	morphometry	diet	tomography	transcripts
17	true	tensor	motivation	resonance	bacteroidetes	feature	plants
18	proportion	mrnas	package	structural	faecal	computed	solanum
19	power	causing	data	healthy	bacteroides	nsclc	wine
20	values	mrna	throughput	maps	microbes	glioblastoma	metabolism

Figure 16: The bag-of-words results for outCitation network with $k = 5$(left) and inCitation network with $k = 7$(right) on dataset II. Each column contains the top twenty representative words in the cluster.

References

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, et al. Construction of the literature graph in semantic scholar. *arXiv preprint arXiv:1805.02262*, 2018.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

Pengsheng Ji, Jiashun Jin, et al. Coauthorship and citation networks for statisticians. *The Annals of Applied Statistics*, 10(4):1779–1812, 2016.

Filippo Radicchi, Santo Fortunato, and Alessandro Vespignani. Citation networks. In *Models of science dynamics*, pages 233–257. Springer, 2012.

Karl Rohe and Muzhe Zeng. Vintage factor analysis with varimax performs statistical inference. *arXiv preprint arXiv:2004.05387*, 2020.

Song Wang and Karl Rohe. Don't mind the (eigen) gap.