Scholarly Network Analysis on Codd's World (The Database Community Publication Graph)

Rutuja Shivraj Pawar, Sepideh Sadat Sobhgol
Under the Guidance of M.Sc. Gabriel Campero Durand
rutuja.pawar@ovgu.de, sepideh.sobhgol@st.ovgu.de, campero@ovgu.de
Team Project - Final Presentation
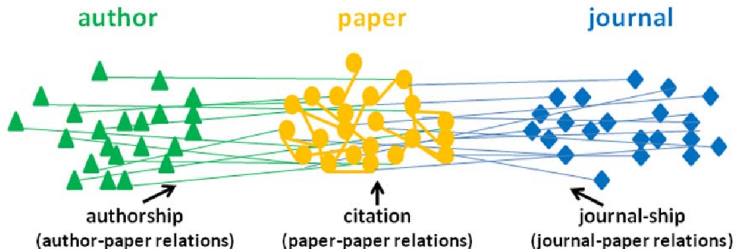October 23, 2018

Agenda

1. **Motivation**

2. **Introduction**

3. **Project Outline**

4. **Research Questions**

5. **Dataset Structure**

6. **Topic Modelling Framework**

7. **Neo4j Framework**

8. **Results**

9. **Conclusion and Scope for Future Work**

Motivation

- Brainchild of the Project: M.Sc. Marcus Pinnecke for the research project ideation and guideline for the formation of the datasets

- Inspiration: Prof. Erhard Rahm's work on Citation Analysis and Affliation Analysis of Database Publications [1] [2] introduced to us by M.Sc. Gabriel Campero Durand

- Our Common Interest: In the field of Data Analytics to draw meaning out of different data collections

Introduction

- What are Scholarly Networks?

- What is Scholarly Network Analysis (SNA)?
  - Network types in SNA
  - Importance of SNA

author   paper   journal

authorship   citation   journal-ship
(author-paper relations)   (paper-paper relations)   (journal-paper relations)

Background of Prof. Erhard Rahm's work on Citation Analysis

- Detailed comparative citation analysis for SIGMOD, VLDB, TODS, VLDB Journal, Sigmod Record over 10 years (1994 - 2003)
- DBLP, Google Scholar and ACM Digital Library
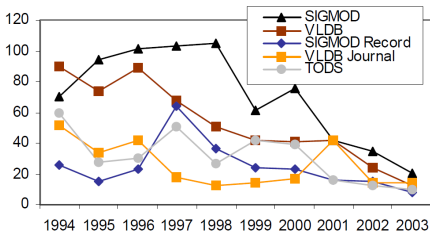- Usefulness of Google Scholar and Data preparation tool iFuice

Figure 1: Average number of citations [1]

Background of Prof. Erhard Rahm's work on Affiliation Analysis

- Detailed author affiliation analysis for ACM SIGMOD, VLDB, ACM TODS, VLDB Journal over 10 years (2000 - 2009)
- ACM Digital Library and SpringerLink
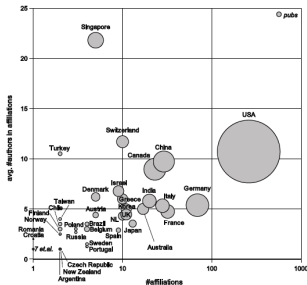- Cross-national co-authorship, degree of collaboration

Figure 2: Country Publications Distribution [2]

Project Outline

Entry Point: Discovering underlying Topics in the network

- SNA on DBLP dataset (Codd's World) [3]
  - Topic Modelling and Visualization
  - Relevance Scoring for Scientific Papers and Authors
  - Dataset Analysis using Neo4j and its query language Cypher
  - Obtaining Insights for the formulated Research Questions
  - Documentation of the obtained Analytical Insights as a Scientific Publication Paper

Research Questions

1. How did **topics evolve in their popularity through time**?
2. Which are the **most cited papers per topic per year**? (*with and without self-citations*)
3. Which is the **most influential paper per topic per year**? *(with and without self-citations)*
4. How **many citations per topic per year**? *(with and without self-citations)*
5. Who is the **most important author per topic**, looking at collaboration only, citation only, and mixed? *(with and without self-citations)*

Dataset Structure

- Specific part of DBLP[3] called Codds World
- Further sub-graphs based on Codd's World
  - The network consisted of four nodes Author, Paper, Venue and Journal.
  - The five types of relationships: authorship, collaboration, belonging to venue, belonging to journal and citation.

Topic Modelling Framework

- Using Python library TOM (TOpic Modeling) [4]
  - Features: Both LDA and NMF Topic Models provided, Optimal Number of Topics estimation, Published Paper [4]
  - Topic Model Selected: Non-negative Matrix Factorization (NMF) [5] [6]
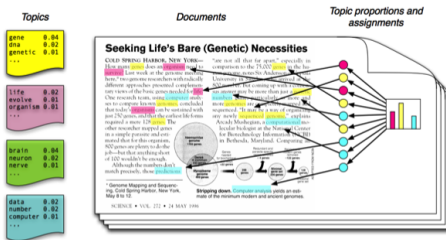 - Optimal Number of Topics Estimation: Estimated as 30 based on the size of the dataset

Figure source: Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77-84.

Neo4j Framework

- Loading of the four nodes and their five types of relationships in Neo4j
- Run of the Relevance Ranking algorithm to calculate ranking scores for Papers and the Authors
- Feeding these scores back into the respective datasets to load and form the final graph in Neo4j
- Cypher Queries to retrieve results for the formulated research questions
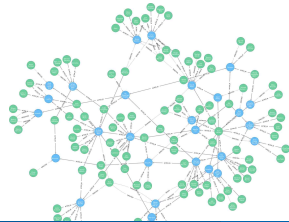
Results RQ1: Discovered Topics

Topic_ID	Topic_Name	Topic_Words
0	NumericalMethods	method proposed method proposed methods based new using method based results estimation
1	Applications	research social study design analysis knowledge technology use human online
2	Networking	network networks nodes routing wireless traffic sensor node protocol neural
3	Optimization	problem problems optimization solution optimal solutions linear set solve function
4	DataMining	data mining data sets data mining sets clustering analysis large database query
5	HCI	performance memory parallel applications high architecture design hardware implementation processor
6	ModelingAndSimulation	model models modeling parameters based process model based simulation proposed model markov
7	Communication	channel signal interference frequency noise channels rate performance multiple error
8	OperatingSystems	system proposed system design based developed paper system performance using monitoring describes
9	CognitiveLearning	learning students machine learning training learn neural student knowledge supervised
10	SemanticWeb	web search semantic query web services pages content services queries documents
11	Algorithms	algorithm algorithms proposed algorithm proposed search based algorithm based clustering new genetic
12	Energy	energy power consumption energy consumption sensor power consumption voltage low efficiency energy efficiency
13	LogicProgramming	language logic languages object semantics programming knowledge programs program semantic
14	ImageProcessing	image images segmentation color 3d object visual resolution regions objects
15	CloudComputing	service services cloud qos computing management quality business resource resources

Figure 3: Topics 0 - 15 given a meaningful name

Results RQ1: Discovered Topics

Topic_ID	Topic_Name	Topic_Words
16	Cryptography	scheme proposed schemes proposed scheme based coding propose signature simulation key
17	ControlTheory	control robot controller robots motion feedback tracking stability loop nonlinear
18	NetworkAnalysis	graph graphs vertices vertex number edge edges set connected tree
19	TimeSeries	time real real time scheduling time series series delay space temporal varying
20	SoftwareEngineering	software development engineering process software development design requirements project tools hardware
21	MachineLearning	features classification feature recognition speech accuracy classifier training detection based
22	VideoProcessing	video motion quality coding frame videos frames content 3d temporal
23	DecisionSupport	fuzzy decision rules sets rule logic clustering set neural controller
24	Testing	test testing fault faults detection tests coverage circuit circuits generation
25	Security	security protocol attacks secure key attack authentication protocols privacy encryption
26	DistributedSystems	agent agents distributed multi communication complex information systems state based
27	BlockCodingAndDecoding	codes code error decoding coding source binary rate length block
28	InformationRetrieval	information retrieval information systems knowledge sources context documents text document available
29	GPSNavigation	user users mobile devices interface interaction device mobile devices location access

Figure 4: Topics 16 - 29 given a meaningful name
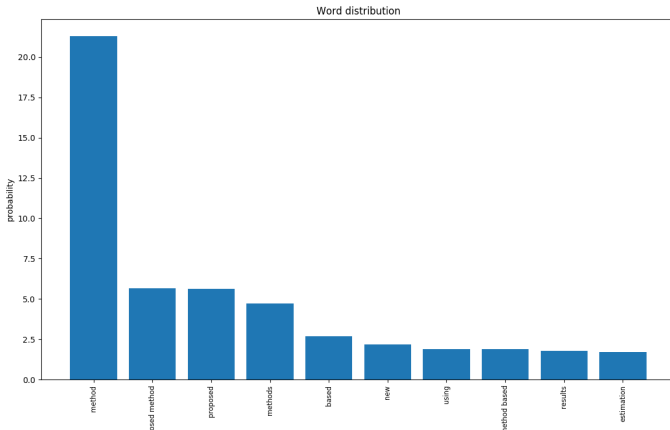
Results RQ1: Word Distribution

Figure 5: Word Distribution for Topic 0

OTTO VON GUERICKE
UNIVERSITÄT
MAGDEBURG

INF

FAKULTÄT FÜR
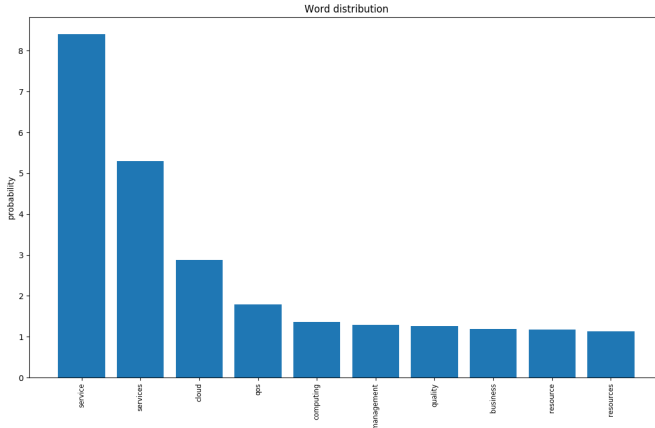INFORMATIK

Results RQ1: Word Distribution

Figure 6: Word Distribution for Topic 15
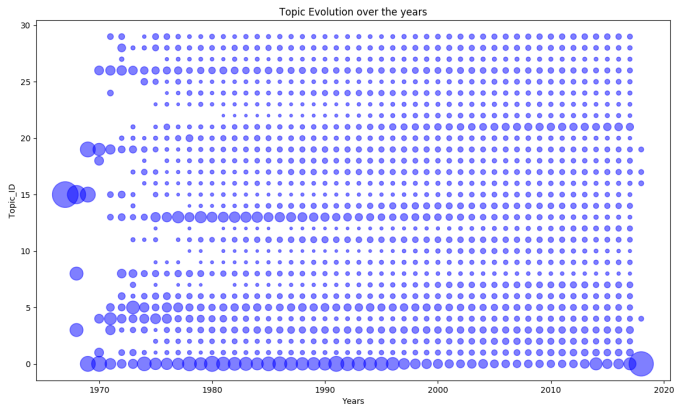
Results RQ1: Topic Evolution

Figure 7: Evolution of Topics through time
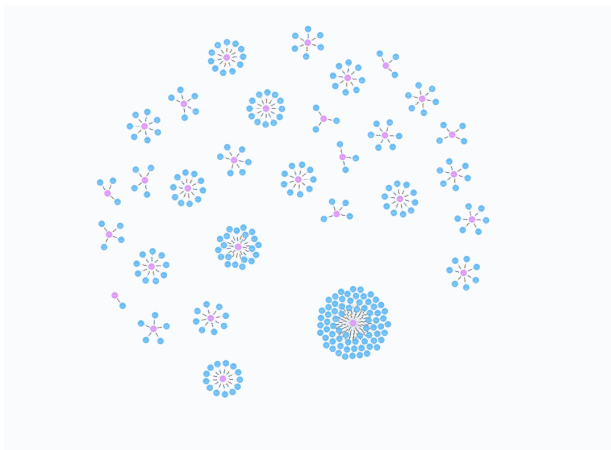
Results RQ1: Topic Clusters

Figure 8: Non-Overlapping Topic Assignments

Observations for RQ1

- Topic 1 (named as Numerical methods) with its highest evolution in the year 2018

- Topic 15 (named as Cloud Computing) with the highest evolution in the year 1965

- Visualization of other topics which have evolved or suddenly disappeared over the years

Results RQ2: Most cited papers in 1970

Tile	TopicName	Count
A Survey of Analytical Time-Sharing Models	NumericalMethods	3
A relational model of data for large shared data banks	DataMining	3
Optimizing the Performance of a Drum-Like Storage	TimeSeries	2
Principles of Optimal Page Replacement	Optimization	1

Table 1: Most cited papers in 1970 with self-citation

Title	TopicName	Count
A Survey of Analytical Time-Sharing Models	NumericalMethods	3
A relational model of data for large shared data banks	DataMining	3
Optimizing the Performance of a Drum-Like Storage	TimeSeries	2

Table 2: Most Cited Papers in 1970 without Self-Citation

Results RQ2: Most cited papers in 2017

Title	TopicName	Count
ImageNet Classification with Deep Convolutional Neural Networks	Testing	736
Caffe: Convolutional Architecture for Fast Feature Embedding	CognitiveLearning	734
LIBSVM: A library for support vector machines	MachineLearning	585
Distinctive Image Features from Scale-Invariant Keypoints	MachineLearning	573
Very Deep Convolutional Networks for Large -Scale Image Recognition	MachineLearning	562
Random Forests	MachineLearning	540
Distributed Representations of Words and Phrases and their Compositionality	CognitiveLearning	490
Histograms of oriented gradients for human detection	MachineLearning	449
Image quality assessment: from error visibility to structural similarity	ImageProcessing	407
Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift	MachineLearning	406

Table 3: Most Cited Papers in 2017 with Self-Citation

FAKULTÄT FÜR
INFORMATIK

OTTO VON GUERICKE
UNIVERSITÄT
MAGDEBURG

INF

Results RQ2: Most cited papers in 2017

Title	TopicName	Count
ImageNet Classification with Deep Convolutional Neural Networks	Testing	736
Caffe: Convolutional Architecture for Fast Feature Embedding	CognitiveLearning	734
LIBSVM: A library for support vector machines	MachineLearning	585
Distinctive Image Features from Scale-Invariant Keypoints	MachineLearning	573
Very Deep Convolutional Networks for Large-Scale Image Recognition	MachineLearning	562
Random Forests	MachineLearning	540
Distributed Representations of Words and Phrases and their Compositionality	CognitiveLearning	490
Histograms of oriented gradients for human detection	MachineLearning	449
Image quality assessment: from error visibility to structural similarity	ImageProcessing	407
Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift	MachineLearning	406

Table 4: Most Cited Papers in 2017 without Self-Citation

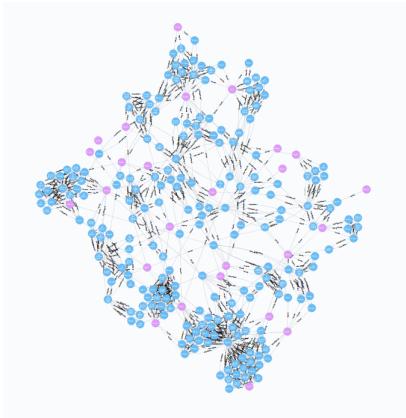# Results RQ2: Top 200 papers without Self-citation 2017

Figure 9: Top 200 papers without Self-Citations in year 2017

Observations for RQ2

- Majority of the top papers with their topics are the same in both the tables of with and without self-citation

- Top papers returned do not achieve their most cited criteria through self-citation

Results RQ3: 25 most Influential papers for all the years

Title	TopicName	Score
A relational model of data for large shared data banks	DataMining	814.4239501953125
Induction of Decision Trees	DistributedSystems	722.2256469726562
Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference	DistributedSystems	544.8001098632812
Snakes: Active Contour Models	MachineLearning	530.5615234375
A theory for multiresolution signal decomposition: the wavelet representation	ImageProcessing	484.8881530761719
A training algorithm for optimal margin classifiers	MachineLearning	431.5468444824219
A robust layered control system for a mobile robot	ControlTheory	362.0459899902344
Highly dynamic Destination-Sequenced Distance-Vector routing (DSDV) for mobile computers	Networking	357.85980224609375
Support-Vector Networks	MachineLearning	357.1878967285156
A learning algorithm for boltzmann machines	HCI	347.3739929199219
The Concept of a Linguistic Variable and its Application to Approximate Reasoning	DecisionSupport	337.3362121582031

Results RQ3: Continued

Title	TopicName	Score
A simple transmit diversity technique for wireless communications	Security	334.5489501953125
MACAW: a media access protocol for wireless LAN's	Security	327.1644592285156
Independent component analysis, a new concept?	DataMining	323.7842712402344
Indexing by Latent Semantic Analysis	SemanticWeb	313.26165771484375
Compliance and Force Control for Computer Controlled Manipulators	ControlTheory	308.8047790527344
Computer architecture: a quantitative approach	HCI	308.58050537109375
Handwritten Digit Recognition with a Back-Propagation Network	Networking	306.81317138671875
A stochastic parts program and noun phrase parser for unrestricted text	MachineLearning	301.9481201171875
Distinctive Image Features from Scale-Invariant Keypoint	MachineLearning	301.7352294921875
LIBSVM: A library for support vector machines	MachineLearning	299.5965270996094
Analysis and simulation of a fair queueing algorithm	Algorithms	296.6663513183594
Supporting real-time applications in an Integrated Services Packet Network: architecture and mechanism	TimeSeries	295.6441345214844
Fast learning in networks of locally-tuned processing units	CognitiveLearning	283.77789306640625
What Size Net Gives Valid Generalization	Testing	281.984619140625

Table 5: 25 most Influential Papers (based on Page Rank score) with Self-Citation for all the years

Results RQ3: 25 most Influential papers for all the years

Title	TopicName	Score
A relational model of data for large shared data banks	DataMining	13669.4931640625
Jobshop-Like Queueing Systems	CloudComputing	5750.12548828125
A model and stack implementation of multiple environments	ControlTheory	5621.76611328125
Toward an understanding of data structures	NetworkAnalysis	5092.0205078125
Procedural embedding of knowledge in planner	Optimization	4842.31005859375
Optimizing the Performance of a Drum-Like Storage	TimeSeries	4267.1123046875
Virtual memory	NumericalMethods	3988.57568359375
New Programming Languages for Artificial Intelligence Research	NumericalMethods	3860.931884765625
Queues with State-Dependent Stochastic Service Rates	CloudComputing	3685.233642578125
Correctness-preserving program transformations	LogicProgramming	3680.9638671875
A universal modular ACTOR formalism for artificial intelligence	NumericalMethods	3617.47900390625
Multiple evaluators in an extensible programming system	LogicProgramming	3111.29833984375

Table 6: 25 most Influential Papers (based on Page Rank score) without Self-Citation for all the years

Results RQ3: Continued

Title	TopicName	Score
Requirements for advanced programming systems for list processing	DistributedSystems	2911.9736328125
Uniqueness of the Gaussian Kernel for Scale-Space Filtering	Communication	2821.81005859375
Scale-space filtering: A new approach to multi-scale description	ImageProcessing	2751.470947265625
Relational Completeness of Data Base Sublanguages	NumericalMethods	2484.283935546875
A Survey of Data Structures for Computer Graphics Systems	DataMining	2473.419677734375
Interference detection among solids and surfaces	Communication	2460.12158203125
Forward Reasoning and Dependency-Directed Backtracking in a System for Computer-Aided Circuit Analysis	OperatingSystems	2418.098388671875
A total standard WIP estimation method for wafer fabrication	Algorithms	2403.2353515625
Higher order approximations for the single server queue with splitting, merging and feedback	DistributedSystems	2403.017578125
Symbolic reasoning among 3-d models and 2-d images	ImageProcessing	2240.699951171875
Abstract data types and software validation	LogicProgramming	2222.979248046875
Induction of Decision Trees	DistributedSystems	2193.14404296875
How to construct random functions	TimeSeries	2152.89111328125

Table 7: 25 most Influential Papers without Self-Citation for all the years

Results RQ3: Most Influential papers 1970

Title	TopicName	Score
A relational model of data for large shared data banks	DataMining	814.4239501953125
Virtual memory	NumericalMethods	151.17889404296875
Toward an understanding of data structures	NetworkAnalysis	27.3781681060791
A schema for describing a relational data base	NumericalMethods	18.157360076904297
Introduction to storage structure definition	NumericalMethods	3.3184258937835693
Time-sharing for OS	TimeSeries	1.6499865055084229
TICKETRON: a successfully operating system without an operating system	DistributedSystems	0.2359350025653839
Swap-Time Considerations in Time-Shared Systems	TimeSeries	0.18187500536441803
A contiuum of time-sharing scheduling algorithms	Applications	0.15000000596046448

Table 8: Most influential papers with self-citation for year 1970

Results RQ3: Most Influential papers 1970

Title	TopicName	Score
A relational model of data for large shared data banks	DataMining	13669.4931640625
Toward an understanding of data structures	NetworkAnalysis	5092.0205078125
Virtual memory	NumericalMethods	3988.57568359375
A schema for describing a relational data base	NumericalMethods	264.00726318359375
Introduction to storage structure definition	NumericalMethods	18.38025665283203
TICKETRON: a successfully operating system without an operating system	DistributedSystems	12.225720405578613
Time-sharing for OS	TimeSeries	5.367920398712158
Swap-Time Considerations in Time-Shared Systems	TimeSeries	0.21375000476837158
A contium of time-sharing scheduling algorithms	Applications	0.15000000596046448

Table 9: 10 most Influential Papers (based on Page Rank score) without Self-Citation 1970

Results RQ3: Most Influential papers 2018

Title	TopicName	Score
Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks	Networking	3.801413059234619
Random Graphs and Complex Networks	NumericalMethods	1.9076725244522095
Minimizing finite sums with the stochastic average gradient	Optimization	1.75485098361969
A Temporal Logic Approach to Binding-Time Analysis	LogicProgramming	1.6572284698486328
On the Linear Convergence of the Alternating Direction Method of Multipliers	Optimization	1.4891154766082764
Order-Optimal Rate of Caching and Coded Multicasting With Random Demands	Security	0.9819459915161133
SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation	HCI	0.7875764966011047
Inventory rebalancing and vehicle routing in bike sharing systems	Optimization	0.7597730159759521
A messy state of the union: taming the composite state machines of TLS	Security	0.6841909885406494
Salient Object Detection: A Discriminative Regional Feature Integration Approach	MachineLearning	0.6486610174179077

Table 10: Most influential papers with self-citation for year 2018

Results RQ3: Most Influential papers 2017

Title	TopicName	Count
Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks	Networking	3.972501039505005
Random Graphs and Complex Networks	NumericalMethods	2.2444255352020264
A Temporal Logic Approach to Binding-Time Analysis	LogicProgramming	2.216188430786133
Minimizing finite sums with the stochastic average gradient	Optimization	1.8653680086135864
On the Linear Convergence of the Alternating Direction Method of Multipliers	Optimization	1.761472463607788
Order-Optimal Rate of Caching and Coded Multicasting With Random Demands	Security	1.1023739576339722
Counting flags in triangle-free digraphs	NetworkAnalysis	0.8560609817504883
SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation	HCI	0.8167909979820251
Inventory rebalancing and vehicle routing in bike sharing systems	Optimization	0.812651515007019
A messy state of the union: taming the composite state machines of TLS	Security	0.6916624903678894

Table 11: 10 most Influential Papers (based on Page Rank score) without Self-Citation 2017
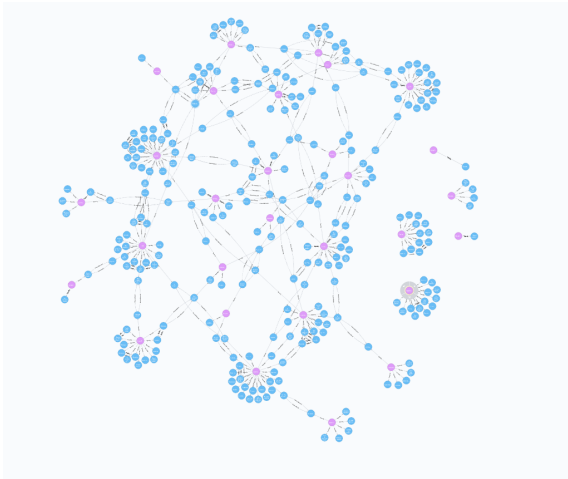
Top Influence per Topic per year

Figure 10: Top Influence per Topic per Year

Observations for RQ3

- The highest Page Rank is indeed associated with the old papers but is not necessarily always true

- As expected, the foundational paper of Edgar Codd on relational databases remains most influential over all the years (with and without self-citation)

- Self-citation makes a difference on the network dynamics but not on the citation count and Page Rank scores depend upon the network structure

- Top Influential papers are cited across topics, because they contribute to different topics, and hence have a high Page Rank

Results RQ4: Citation Counts per year
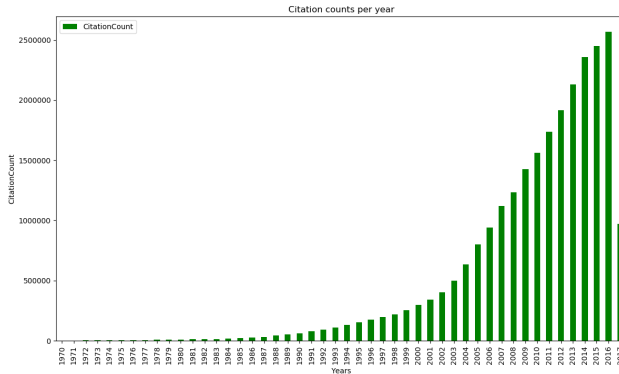
Figure 11: Total citation count per year

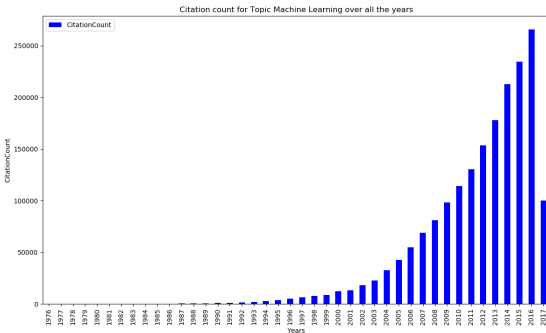# Results RQ4: Citation Count for Topic Machine Learning over the years

Figure 12: Citation Count for Topic Machine Learning over the years

Observations for RQ4

- An increasing trend for the selected topic Machine Learning

- No significant downtrend observed for any other topic

Results RQ5: 25 most Influential Authors based on Author Rank

AuthorName	Score
Wei Wan.	215.6251220703125
Wei Zhan.	146.0390625
Wei L.	143.38377380371094
Lei Wan.	142.0850372314453
Yang Li.	137.28176879882812
Lei Zhan.	128.57749938964844
Wei Che.	121.05441284179688
Jun Wan.	117.1329345703125
Wei Li.	115.75605010986328
Xin L.	108.65744018554688
Yan Zhan.	106.84562683105469
Li Zhan.	104.4130859375
Jun Zhan.	104.29021453857422
Yang Yan.	99.6176986694336
Jing Wan.	98.92251586914062

AuthorName	Score
Yu Zhan.	97.22993469238281
Xin Wan.	95.66221618652344
Li L.	94.99837493896484
Jing L.	92.20679473876953
Jie Zhan.	90.3498764038086
Jun L.	89.5878677368164
Yu Wan.	89.09492492675781
Hui L.	88.64311218261719
Yan L.	85.41546630859375
Yang L.	83.77407836914062

Table 12: Author Rank on all Topics (25 most Influential Authors)

Results RQ5: 25 most Influential Authors based on combination of Author Rank and Page Rank (without Self-Citation)

AuthorName	Score
E. F. Cod.	18399.134765625
Daniel G. Bobro.	14275.8740234375
Carl Hewit.	12347.6787109375
Ben Wegbrei.	9271.6328125
Andrew P. Witki.	7430.35205078125
Peter J. Dennin.	7198.57421875
Robert Endre Tarja.	6088.55419921875
Peter Boehler Bisho.	5915.26806640625
Richard Steige.	5915.26806640625
James R. Jackso.	5750.12548828125
Jay Earle.	5562.37841796875
H. T. Kun.	4987.71875
Rodney A. Brook.	4939.78271484375
Geoffrey E. Hinto.	4618.77734375
Joseph Abat.	4354.34765625
Richard P. Bren.	4338.82666015625
David R. Musse.	4316.86474609375

AuthorName	Score
Harvey Dubne.	4279.337890625
Ellis Horowit.	4257.8330078125
Robert L. Merce.	4257.7353515625
Larry S. Davi.	3981.80810546875
Bertram Raphae.	3860.931884765625
Susan L. Gerhar.	3806.7587890625
Oded Goldreic.	3787.566650390625
David Haussle.	3755.4345703125

Table 13: 25 most Influential Authors without Self-Citation (Combination of Author Rank and Page Rank)

Results RQ5: 25 most Influential Authors on all Topics with Self-Citation based on Page Rank

AuthorName	Score
Scott Shenke.	2323.373779296875
Demetri Terzopoulo.	1693.306396484375
Robert L. Merce.	1608.9576416015625
Geoffrey E. Hinto.	1563.3994140625
Hari Balakrishna.	1534.552978515625
Rakesh Agrawa.	1505.221923828125
Vladimir Vapni.	1460.100830078125
Andrew P. Witki.	1459.3623046875
Deborah Estri.	1458.5408935546875
Lixia Zhan.	1445.181884765625
Alex Pentlan.	1413.5933837890625
E. F. Cod.	1409.331787109375
David E. Culle.	1386.3238525390625
Anil K. Jai.	1343.75537109375
David Haussle.	1284.4566650390625
Robert E. Schapir.	1277.0406494140625
Frederick Jeline.	1244.083740234375
Ian T. Foste.	1220.2354736328125

AuthorName	Score
Judea Pear.	1195.59033203125
Rodney A. Brook.	1189.1766357421875
Takeo Kanad.	1180.5196533203125
Bernhard Schlkop.	1172.0855712890625
Sally Floy.	1149.4163818359375
Michael I. Jorda.	1136.1651611328125
Michael Stonebrake.	1126.252685546875

Table 14: 25 most Influential Authors on all Topics with Self-Citation

Results RQ5: 25 most Influential Authors on all Topics without Self-Citation based on Page Rank

AuthorName	Score
E. F. Codd.	18399.134765625
Daniel G. Bobro.	14275.8740234375
Carl Hewit.	12347.6787109375
Ben Wegbrei.	9271.6328125
Andrew P. Witki.	7430.35205078125
Rakesh Agrawa.	1505.221923828125
Vladimir Vapni.	1460.100830078125
Andrew P. Witki.	1459.3623046875
Peter J. Dennin.	7198.57421875
Robert Endre Tarja.	6088.55419921875
Peter Boehler Bisho.	5915.26806640625
Richard Steige.	5915.26806640625
James R. Jackso.	5750.12548828125
Jay Earle.	5562.37841796875
H. T. Kun.	4987.71875
Rodney A. Brook.	4939.78271484375
Geoffrey E. Hinto.	4618.77734375

AuthorName	Score
Joseph Abat.	4354.34765625
Richard P. Bren.	4338.82666015625
David R. Musse.	4316.86474609375
Harvey Dubne.	4279.337890625
Ellis Horowit.	4257.8330078125
Robert L. Merce.	4257.7353515625
Larry S. David.	3981.80810546875
Bertram Raphae.	3860.931884765625
Susan L. Gerhar.	3806.7587890625
Oded Goldreic.	3787.566650390625
David Haussle.	3755.4345703125

Table 15: 25 most Influential Authors on all Topics without Self-Citation

Results RQ5: Top Authors with Self-Citation for Data Mining (Combination of Author Rank and Page Rank)

AuthorName	Score
Scott Shenke.	2323.373779296875
Demetri Terzopoulo.	1693.306396484375
Geoffrey E. Hinto.	1563.3994140625
Hari Balakrishna.	1534.552978515625
Rakesh Agrawa.	1505.221923828125
Vladimir Vapni.	1460.100830078125
Deborah Estri.	1458.5408935546875
Lixia Zhan.	1445.181884765625
Alex Pentlan.	1413.5933837890625
E. F. Cod.	1409.331787109375
David E. Culle.	1386.3238525390625
Anil K. Jai.	1343.75537109375
David Haussle.	1284.4566650390625
Robert E. Schapir.	1277.0406494140625
Ian T. Foste.	1220.2354736328125
Judea Pear.	1195.59033203125
Takeo Kanad.	1180.5196533203125
Bernhard Schlkop.	1172.0855712890625

AuthorName	Score
Michael I. Jorda.	1136.1651611328125
Jitendra Mali.	1117.4991455078125
Alan J. Demer.	1100.455322265625
Christos Faloutso.	1052.6434326171875
David R. Karge.	1042.8453369140625
Robert Morri.	1023.3004150390625

Table 16: Top Authors with Self-Citation for Data Mining

Results RQ5: Top Authors without Self-Citation for Data Mining (Combination of Author Rank and Page Rank)

AuthorName	Score
E. F. Cod.	18399.134765625
Daniel G. Bobro..	14275.8740234375
Carl Hewit.	12347.6787109375
Ben Wegbrei.	9271.6328125
Peter J. Dennin.	7198.57421875
Robert Endre Tarja.	6088.55419921875
Peter Boehler Bisho.	5915.26806640625
Richard Steige.	5915.26806640625
H. T. Kun.	4987.71875
Geoffrey E. Hinto.	4618.77734375
David R. Musse.	4316.86474609375
Ellis Horowit.	4257.8330078125
Larry S. Davi.	3981.80810546875
David Haussle.	3755.4345703125
Michael Stonebrake.	3747.854248046875
Don Chamberli.	3745.54296875
Linda G. Shapir.	3681.767333984375

AuthorName	Score
Silvio Mical.	3540.778076171875
Jim Gra.	3488.064453125
Raymond A. Lori.	3453.755615234375
Zohar Mann.	3450.918701171875
Terrence J. Sejnowsk.	3410.999755859375
Demetri Terzopoulo.	3395.916748046875
Scott Shenke.	3378.406494140625
Azriel Rosenfel.	3370.689697265625

Table 17: Top Authors without Self-Citation for Data Mining)

Results RQ5: Collaboration/ Co-authorship network of Prof. Gunter Saake
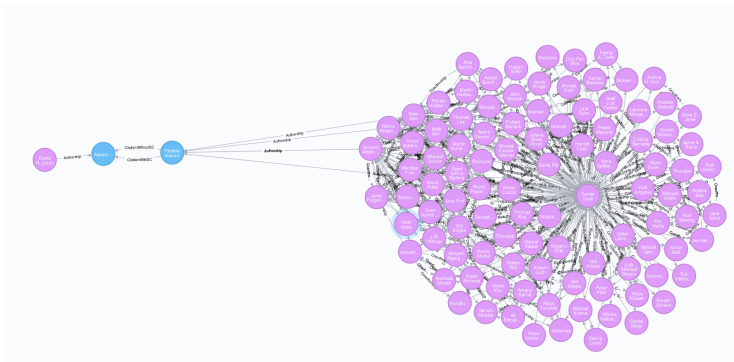
Figure 13: Collaboration/Co-Authorship Network of Prof. Gunter Saake

Observations for RQ5

- Author Rank on all topics indicates how spread is the collaboration with other authors

- Combination of Author Rank and Page Rank does lead to different results in the ranking of authors in comparison to Page Rank alone mostly because the collaboration network had very small weights

- Most collaborative authors publish papers in all the topics

Conclusion and Scope for Future Work

- Analytical Insights obtained related to various aspects of the analyzed database scholarly network
- Topic Evolution, Scientific impact of a paper, Paper acceptance trend, Popularity of research on a topic, Author's impact in the research community
- Better research understanding of the database scientific publication network and facilitate data-driven research decisions

Conclusion and Scope for Future Work

- Increasing Topic Accuracy based on stability analysis by Greene Matrix [7]
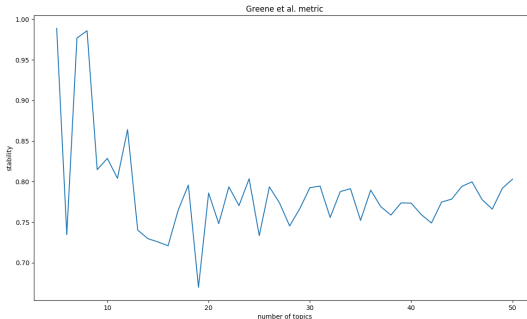
Figure 14: Greene Matrix plotted for the input dataset

Conclusion and Scope for Future Work

- Including information regarding distribution of the papers having high citation counts to find accurate trends. Using Power Law analysis [8] to drill down into papers

- Elimination of the dataset limitation of duplicate author names which may result in vagueness of results for ranking authors

- Formulation of different Research questions, Extension of Research to other fields *to follow a divide and conquer approach to uncover insights from the ultimate Big Scholarly Data*

Thank You all for your attention!
Any Questions?

📄 E. Rahm and A. Thor, "Citation analysis of database publications," *ACM Sigmod Record*, vol. 34, no. 4, pp. 48–53, 2005.

📄 D. Aumüller and E. Rahm, "Affiliation analysis of database publications," *ACM SIGMOD Record*, vol. 40, no. 1, pp. 26–31, 2011.

📄 M. Ley, "The dblp computer science bibliography: Evolution, research issues, perspectives," in *International symposium on string processing and information retrieval*. Springer, 2002, pp. 1–10.

📄 A. Guille and E.-P. Soriano-Morales, "Tom: A library for topic modeling and browsing." in *EGC*, 2016, pp. 451–456.

📄 M. W. Berry and M. Browne, "Email surveillance using non-negative matrix factorization," *Computational &*

Mathematical Organization Theory, vol. 11, no. 3, pp. 249–264, 2005.

📄 D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, p. 788, 1999.

📄 D. Greene, D. OCallaghan, and P. Cunningham, "How many topics? stability analysis for topic models," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, 2014, pp. 498–513.

📄 A. Clauset, C. R. Shalizi, and M. E. Newman, "Power-law distributions in empirical data," *SIAM review*, vol. 51, no. 4, pp. 661–703, 2009.

📄 Scholarly Network. [Image] *A heterogeneous scholarly network*. https://www.semanticscholar.org/paper/This-is-the-

preliminary-JASIST-submission-PRank-
%3A−Yan−Ding/92507ec471d207b65c99c5d517bd08fc5
0671ced /figure/0

📄 Neo4j Logo. [Image] *Neo4j Logo*. https://neo4j.com/

📄 Neo4j Graph. [Image] *Neo4j Graph*.
https://smalltalk.connpass.com/event/64380/