Selecting Most Influential Topics In The Future

Zhao Jinghao

5140309102

CONTENTS

Abstract

PART ONE

Topic Map

PART TWO

Features Extraction

PART THREE

Prediction

PART FOUR

Conclusion

PART FIVE

References

PART SIX

Abstract

PART ONE

Abstract

K-core Analysis of Scholarly Network

Visualization of basic structure of the CS field and show the relationship between different topics. Including topic clustering, k-core analysis, and heat representing.

Topic Factor Extraction

Extract factors that can influence the future development or the factor that can show the present state of a topic. Determine how these factors influence the growth rate of the topic.

Topic Scale Prediction

The goal is to regard the scale prediction as a regression problem. Given the factor matrix M of topic T at time t, the problem is to predict the paper number N, which means the size and scale of this topic, at the time t + Δt .

Topic Map

PART TWO

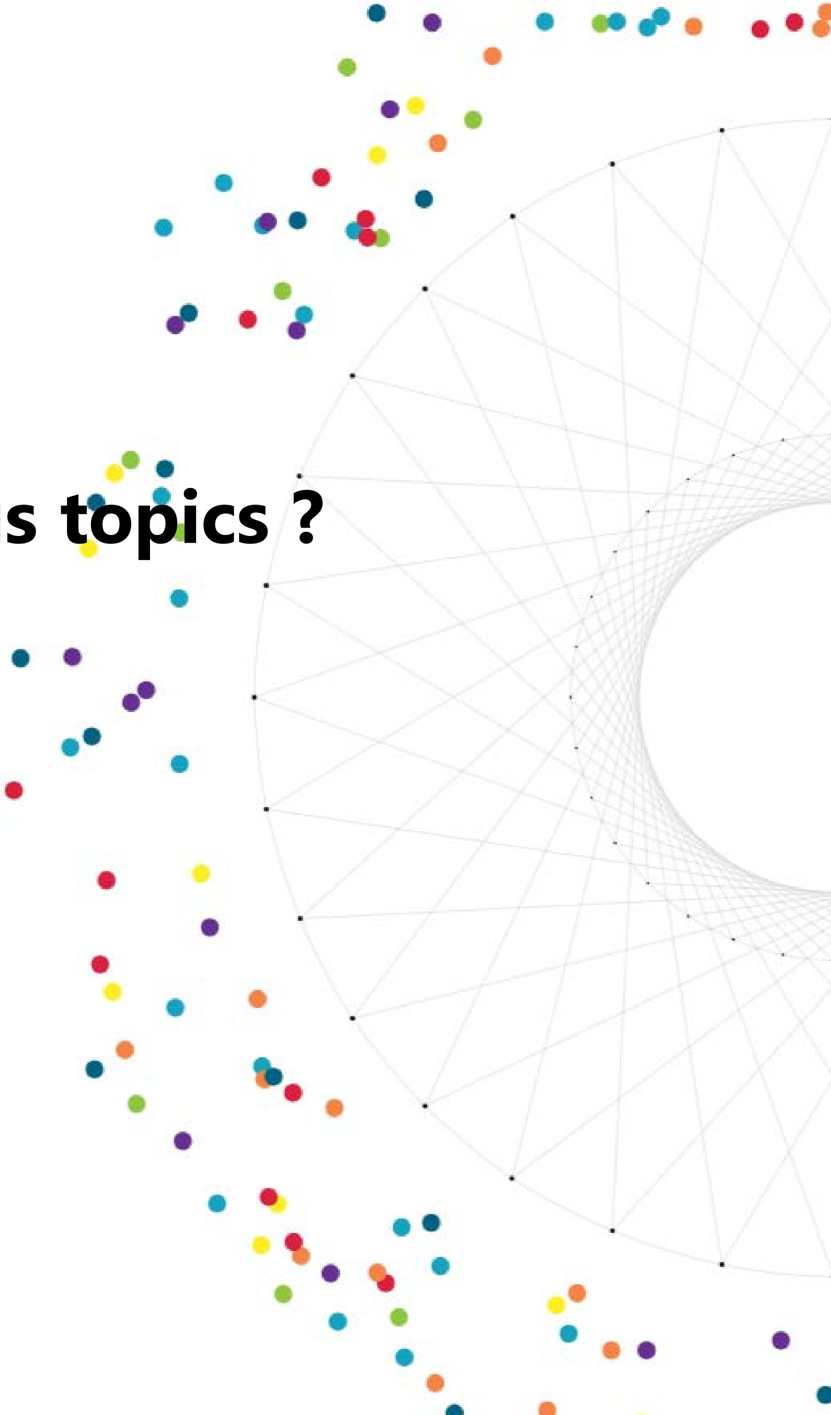
Topic Map

What does the computer domain contain ?

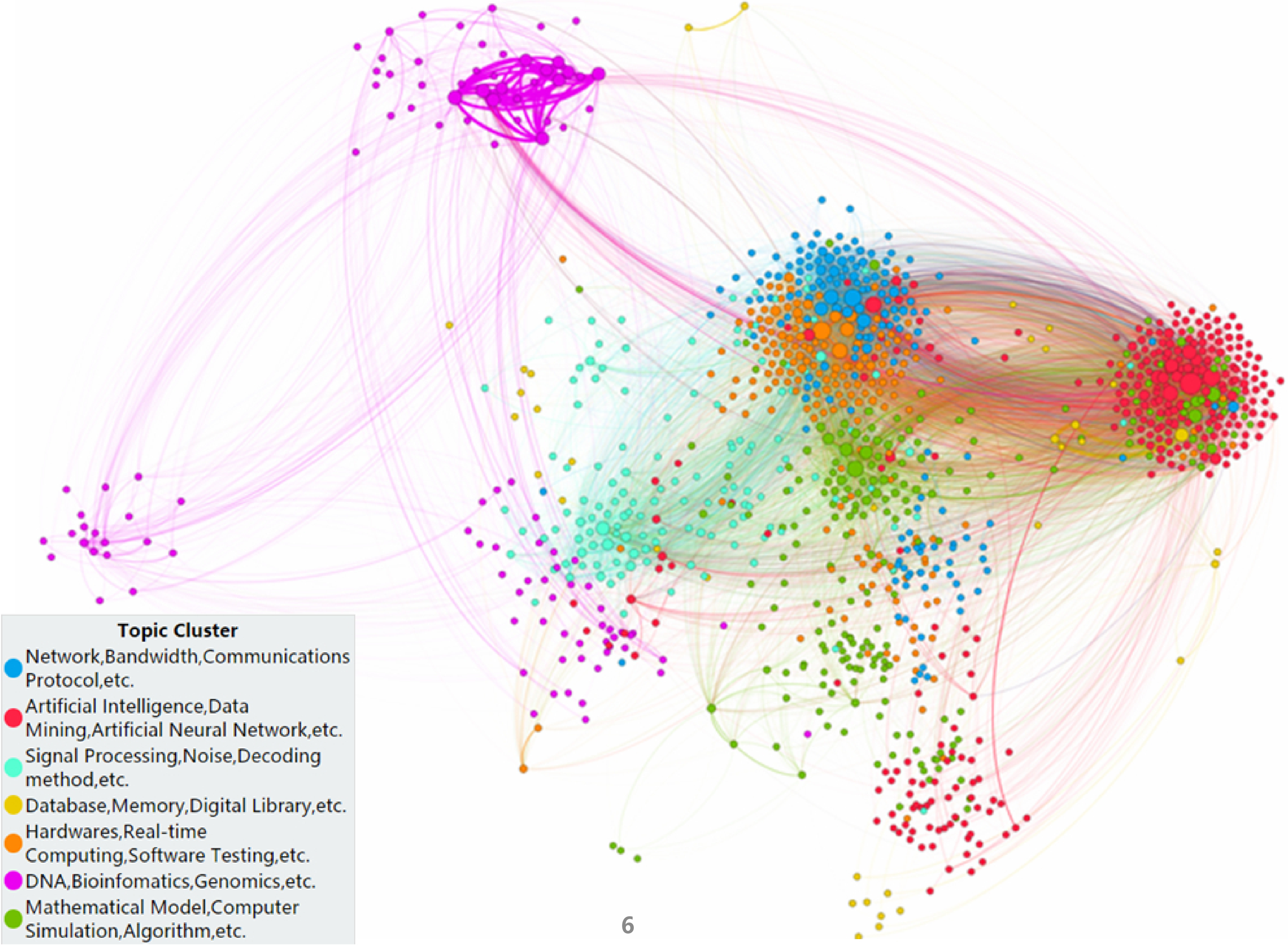
What is the relationship between the various topics ?

What is the basis of the computer field ?

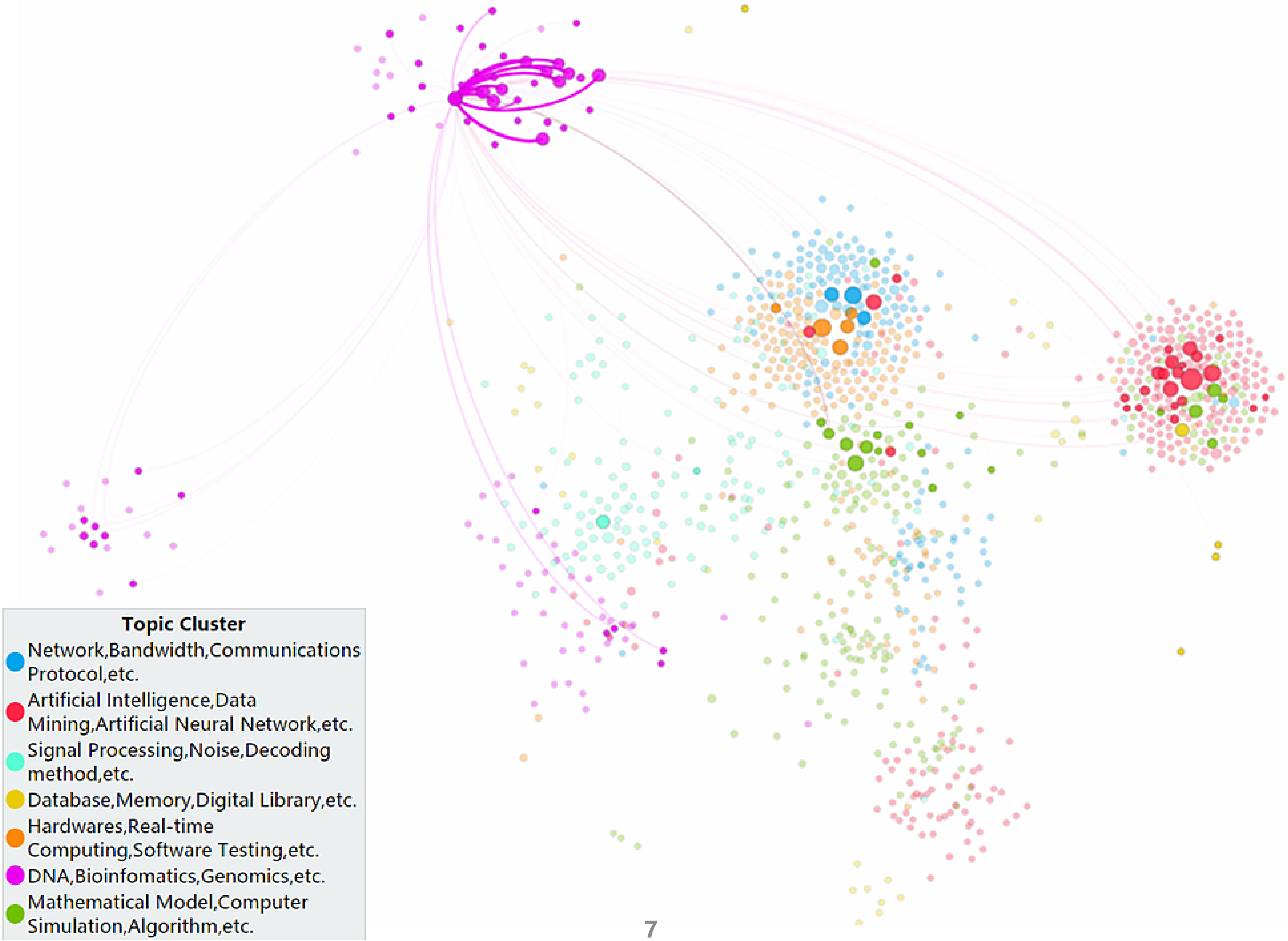
How various topics influence each other ?



Topic Map

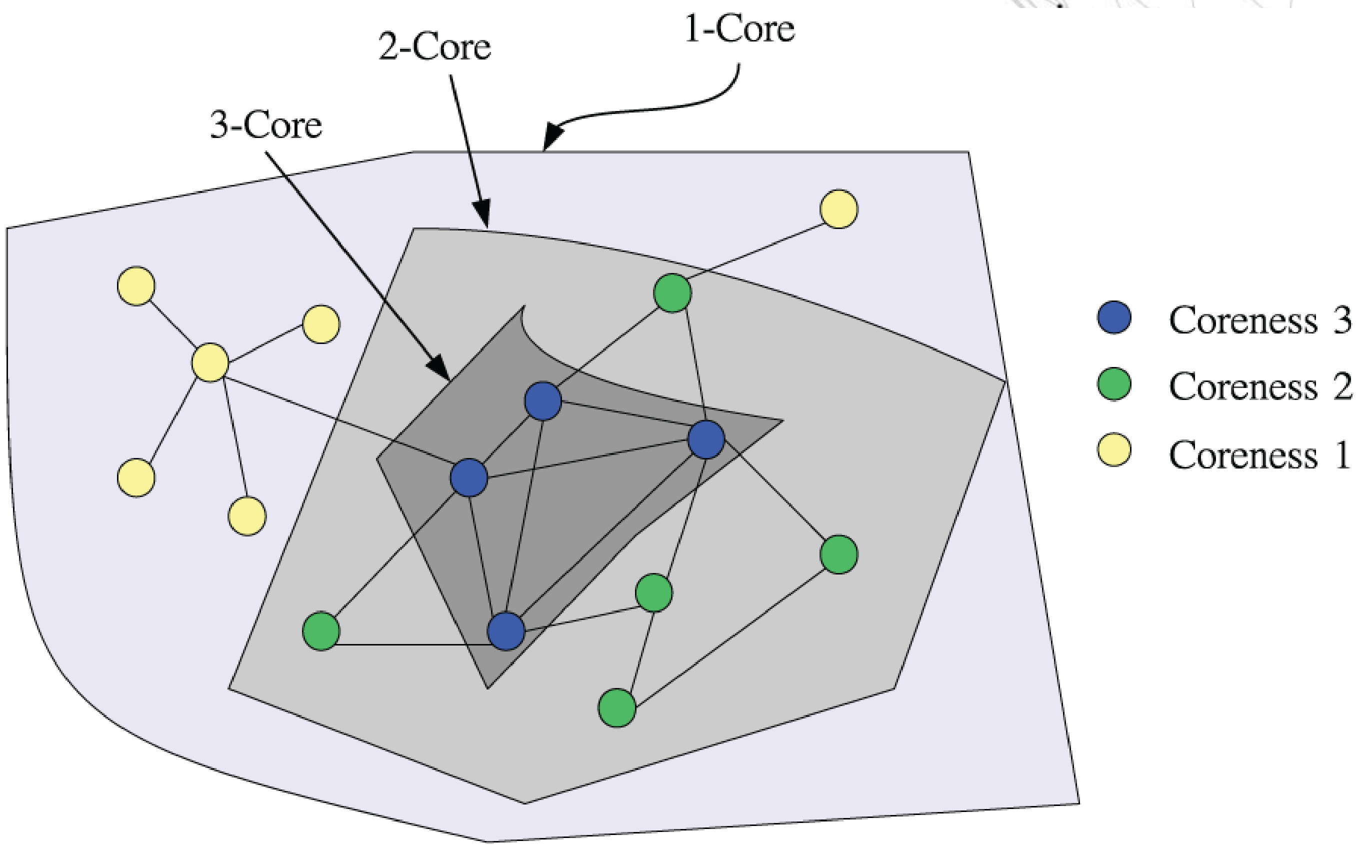

Topic Cluster
● Network,Bandwidth,Communications Protocol,etc.
● Artificial Intelligence,Data Mining,Artificial Neural Network,etc.
● Signal Processing,Noise,Decoding method,etc.
● Database,Memory,Digital Library,etc.
● Hardwares,Real-time Computing,Software Testing,etc.
● DNA,Bioinfomatics,Genomics,etc.
● Mathematical Model,Computer Simulation,Algorithm,etc.

Topic Map

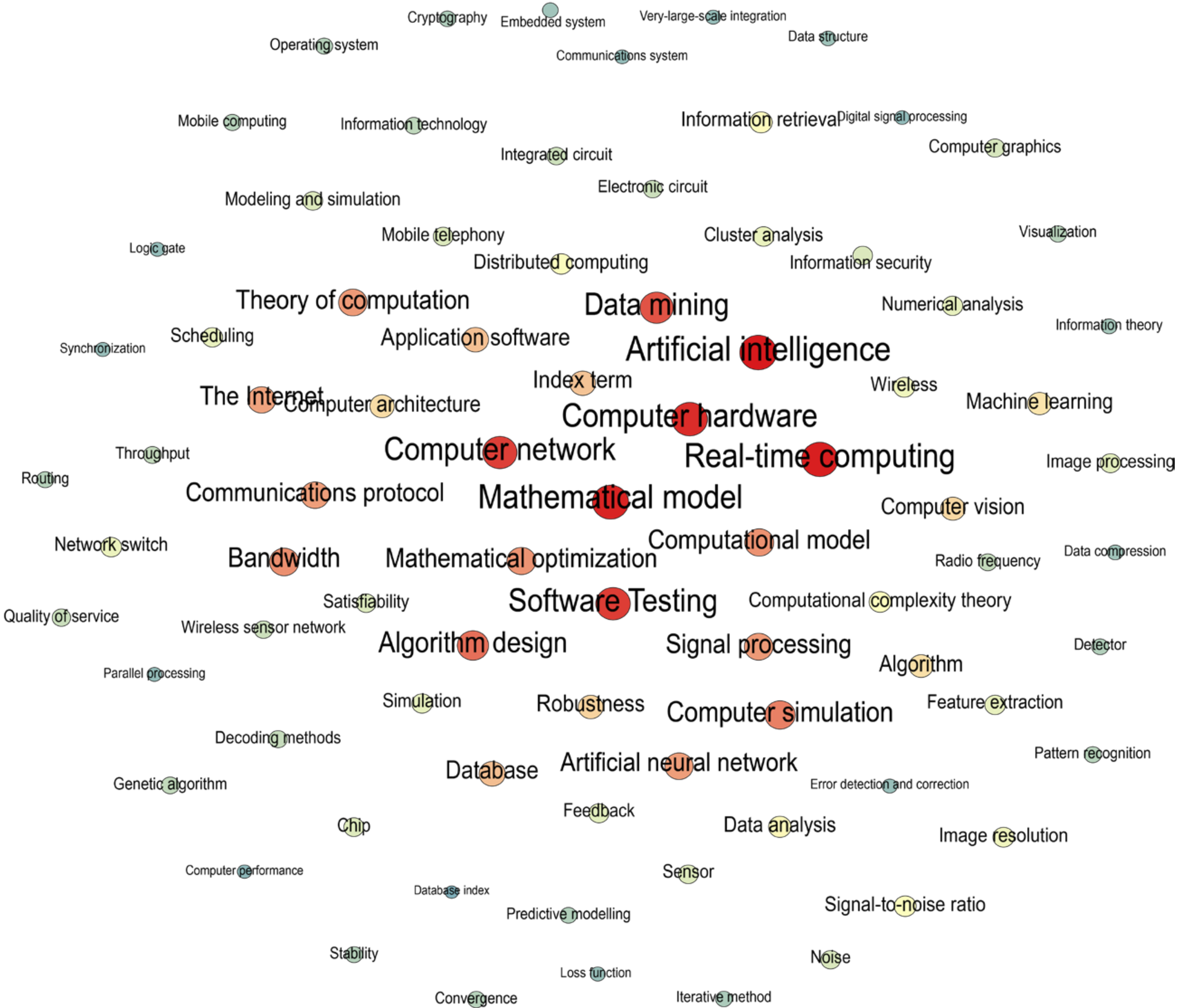

Topic Cluster
● Network,Bandwidth,Communications Protocol,etc.
● Artificial Intelligence,Data Mining,Artificial Neural Network,etc.
● Signal Processing,Noise,Decoding method,etc.
● Database,Memory,Digital Library,etc.
● Hardwares,Real-time Computing,Software Testing,etc.
● DNA,Bioinfomatics,Genomics,etc.
● Mathematical Model,Computer Simulation,Algorithm,etc.

K-core Analysis of Scholarly Network

A **k-core** is the maximal subgraph where all vertices have degree at least k.



Topic Map


Cryptography
Embedded system
Very-large-scale integration
Operating system
Data structure
Communications system
Mobile computing
Information technology
Information retrieval
Digital signal processing
Integrated circuit
Computer graphics
Electronic circuit
Modeling and simulation
Logic gate
Mobile telephony
Cluster analysis
Visualization
Distributed computing
Information security
Theory of computation
Data mining
Numerical analysis
Scheduling
Application software
Information theory
Synchronization
Artificial intelligence
Index term
Wireless
The Internet
Computer architecture
Machine learning
Computer hardware
Throughput
Computer network
Real-time computing
Image processing
Routing
Communications protocol
Mathematical model
Computer vision
Computational model
Network switch
Bandwidth
Mathematical optimization
Radio frequency
Data compression
Satisfiability
Software Testing
Computational complexity theory
Quality of service
Wireless sensor network
Algorithm design
Signal processing
Detector
Algorithm
Parallel processing
Simulation
Robustness
Computer simulation
Feature extraction
Decoding methods
Pattern recognition
Database
Artificial neural network
Genetic algorithm
Error detection and correction
Feedback
Chip
Data analysis
Image resolution
Computer performance
Sensor
Database index
Predictive modelling
Signal-to-noise ratio
Stability
Noise
Loss function
Convergence
Iterative method

Features Extraction

PART THREE

Features Extraction

Paper Factor

- Paper-num
- Citation-ave
- Citation-max

Author Factor

- Author-num
- Author-hindex-ave
- Author-hindex-max
- Author-hindex-var

Growth Factor

- Increase-num
- Increase-num-ave
- Increase-num-max

Venue Factor

- Venue-num
- Venue-distinct-num
- Venue-index-ave

Interaction Factor

- interaction-growthnum-ave
- interaction-growthnum-max

Features Extraction

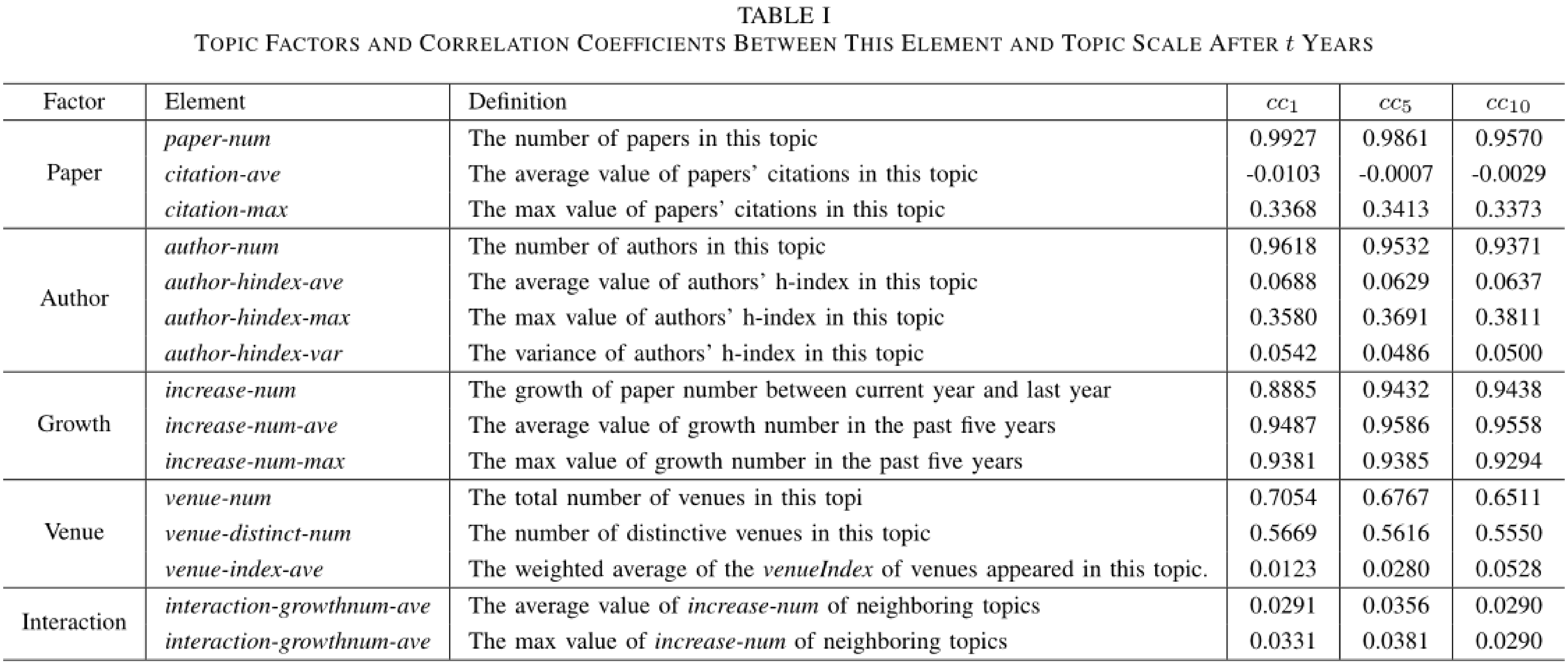


TABLE I
TOPIC FACTORS AND CORRELATION COEFFICIENTS BETWEEN THIS ELEMENT AND TOPIC SCALE AFTER t YEARS

Factor	Element	Definition	cc_1	cc_5	cc_{10}
Paper	*paper-num*	The number of papers in this topic	0.9927	0.9861	0.9570
	citation-ave	The average value of papers' citations in this topic	-0.0103	-0.0007	-0.0029
	citation-max	The max value of papers' citations in this topic	0.3368	0.3413	0.3373
Author	*author-num*	The number of authors in this topic	0.9618	0.9532	0.9371
	author-hindex-ave	The average value of authors' h-index in this topic	0.0688	0.0629	0.0637
	author-hindex-max	The max value of authors' h-index in this topic	0.3580	0.3691	0.3811
	author-hindex-var	The variance of authors' h-index in this topic	0.0542	0.0486	0.0500
Growth	*increase-num*	The growth of paper number between current year and last year	0.8885	0.9432	0.9438
	increase-num-ave	The average value of growth number in the past five years	0.9487	0.9586	0.9558
	increase-num-max	The max value of growth number in the past five years	0.9381	0.9385	0.9294
Venue	*venue-num*	The total number of venues in this topi	0.7054	0.6767	0.6511
	venue-distinct-num	The number of distinctive venues in this topic	0.5669	0.5616	0.5550
	venue-index-ave	The weighted average of the *venueIndex* of venues appeared in this topic.	0.0123	0.0280	0.0528
Interaction	*interaction-growthnum-ave*	The average value of *increase-num* of neighboring topics	0.0291	0.0356	0.0290
	interaction-growthnum-ave	The max value of *increase-num* of neighboring topics	0.0331	0.0381	0.0290

Time Serialization

Time serialization to each factor of **12243 topics**

From **1950 to 2015**

Containing

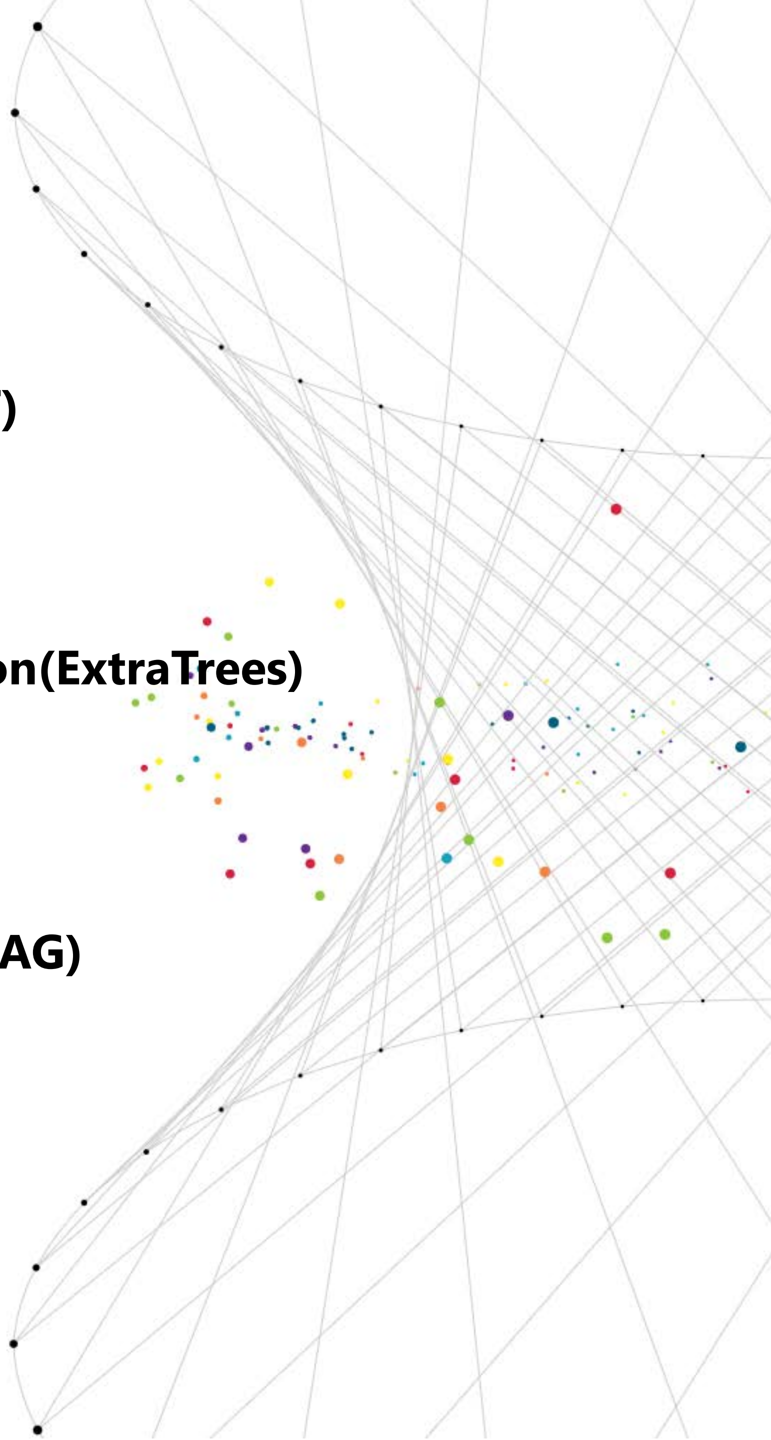
more than **14.4 million** authors

more than **30 million** papers

Prediction

PART FOUR

Models



Linear regression(LR)

Decision Tree Regression(DT)

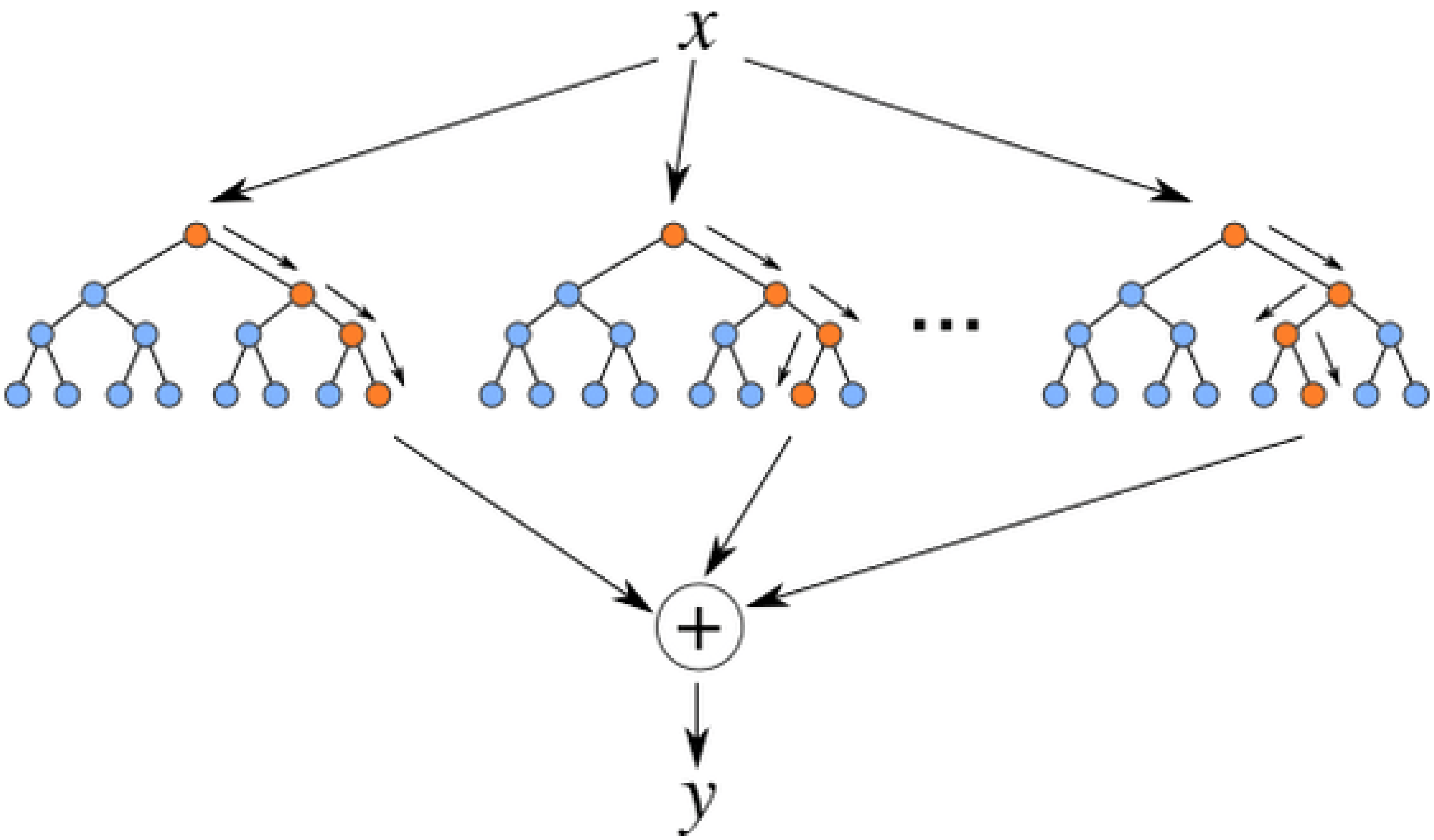
Random Forest Regression(RF)

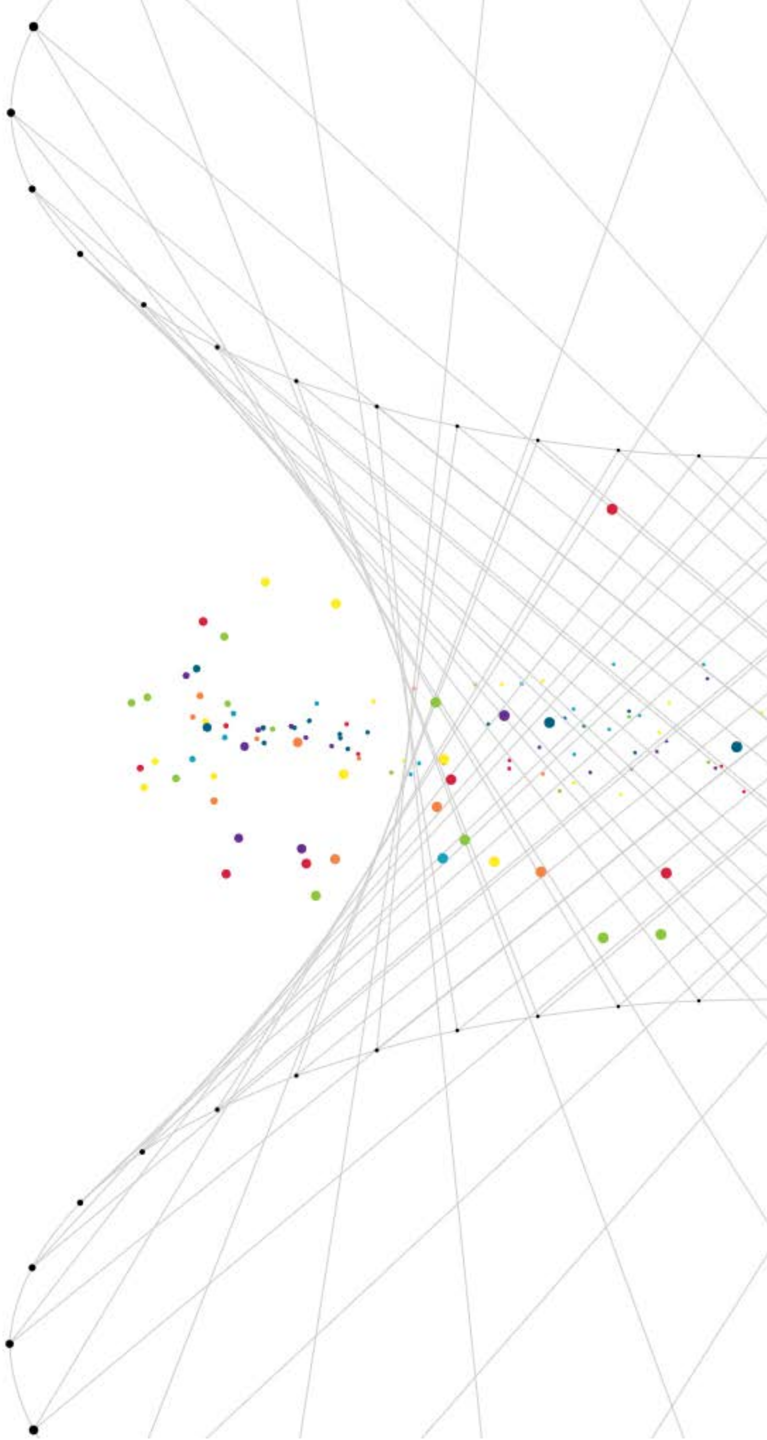
Extremely Randomized Trees Regression(ExtraTrees)

Gradient Boosting Regression(GBDT)

bagged decision trees(BAG)

Random Forest





Predictive Performance

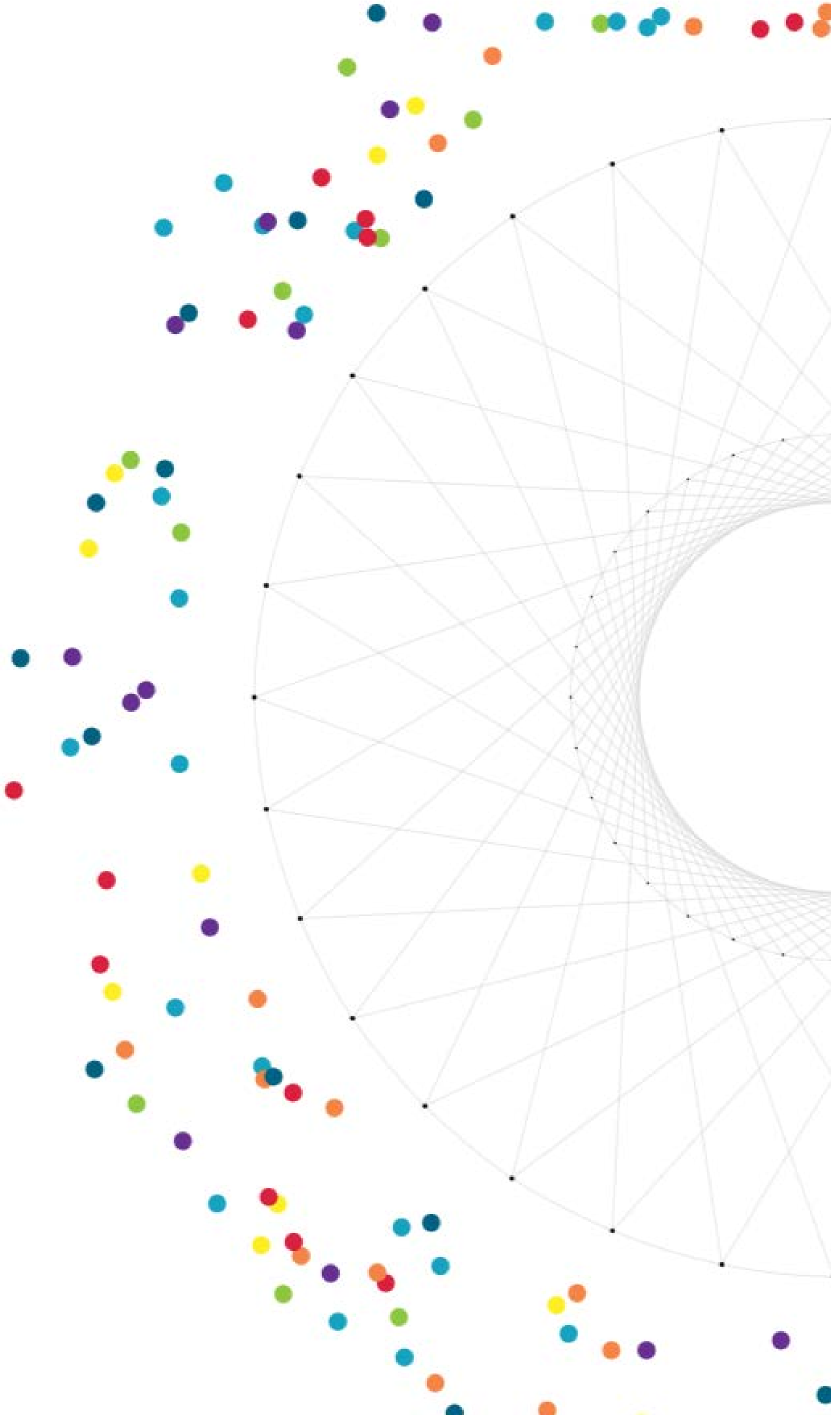
1. Coefficient of Determination (R^2)

$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2, \quad SS_{\text{res}} = \sum_i (y_i - f_i)^2 = \sum_i e_i^2$$

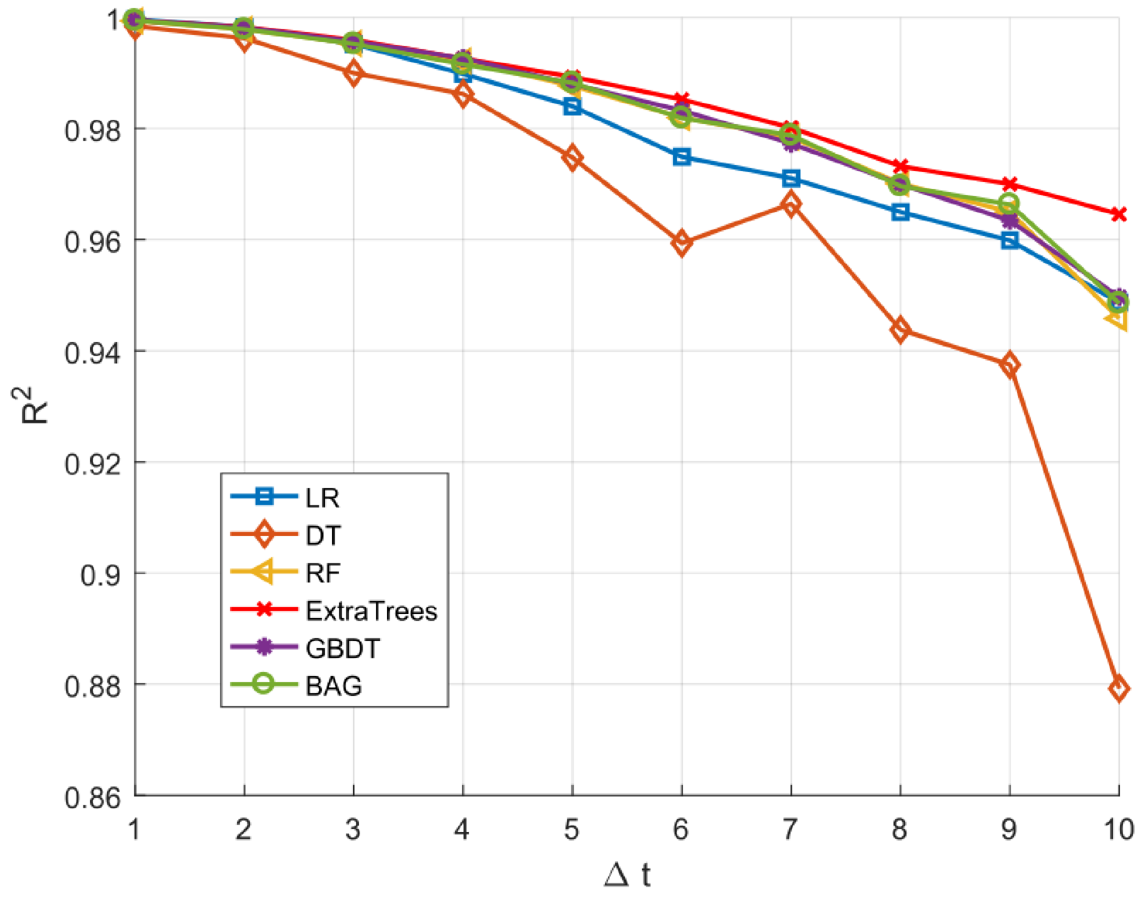
$$R^2 \equiv 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

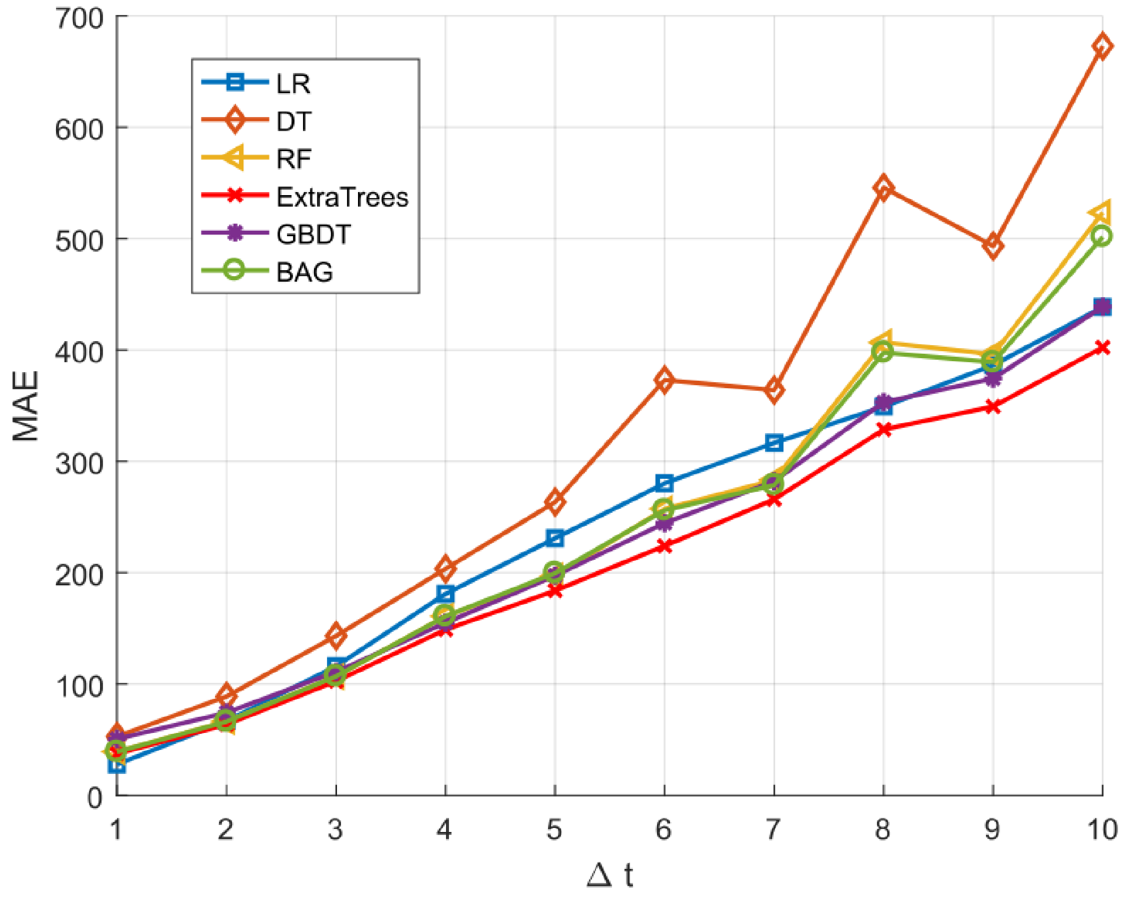
2. Mean Absolute Error (MAE)

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n}$$



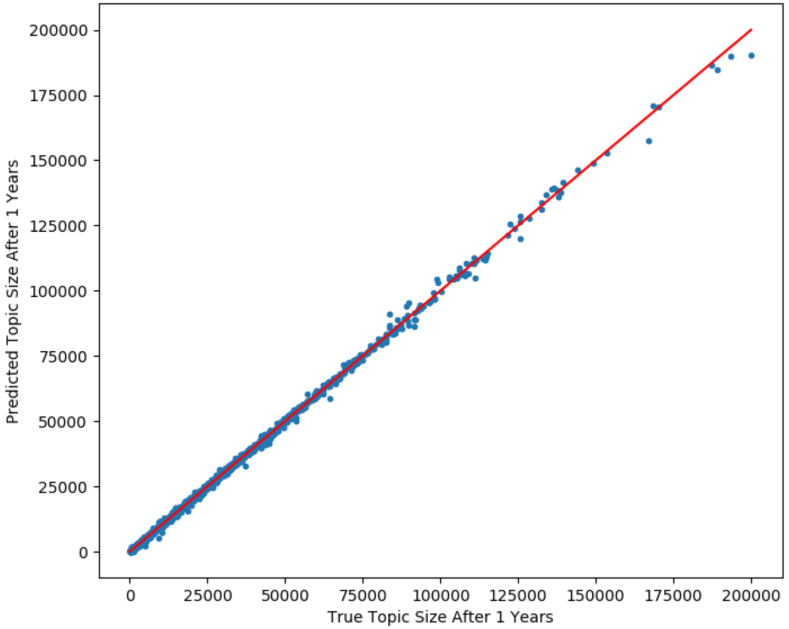
Prediction


R²
Δ t
1
0.98
0.96
0.94
0.92
0.9
0.88
0.86
1
2
3
4
5
6
7
8
9
10
LR
DT
RF
ExtraTrees
GBDT
BAG

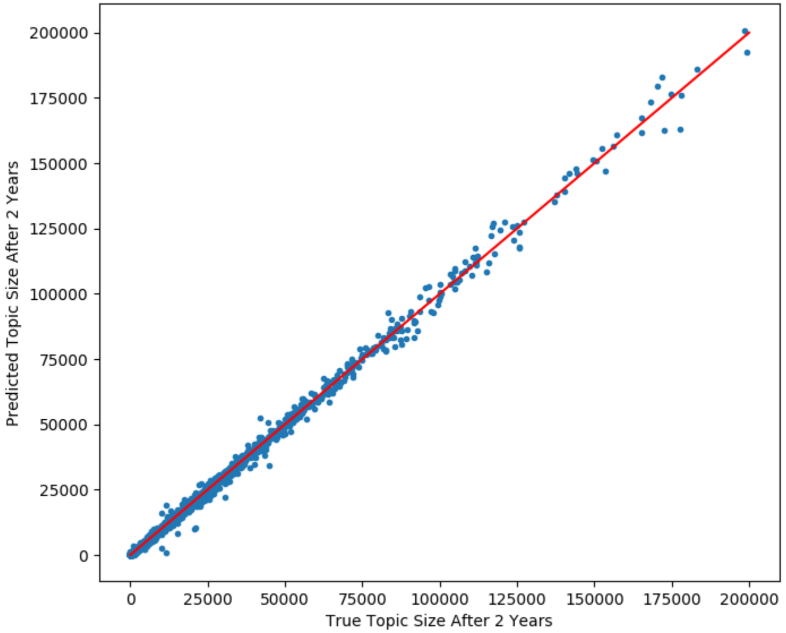

MAE
Δ t
700
600
500
400
300
200
100
0
1
2
3
4
5
6
7
8
9
10
LR
DT
RF
ExtraTrees
GBDT
BAG

Prediction

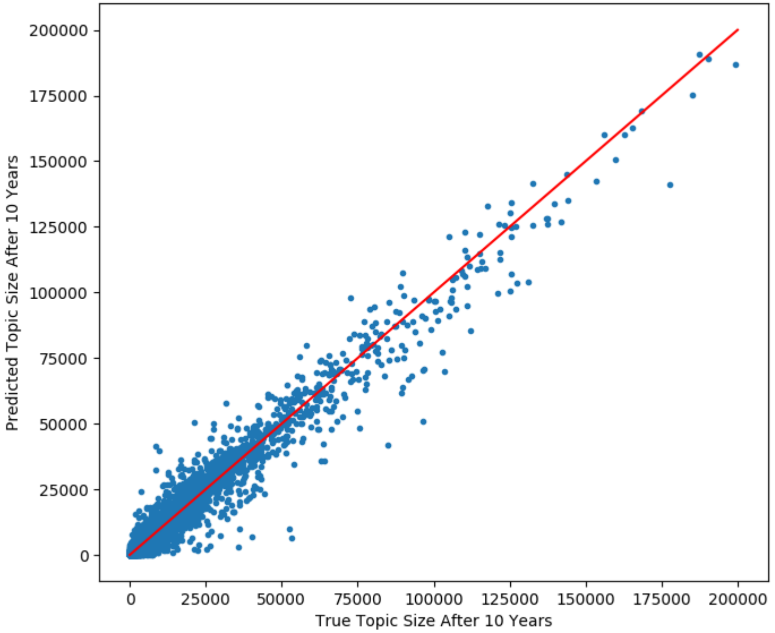
Δt=1



Δt=5



Δt=10



Prediction

Top 100 Hot Topic in 5 years

Exome sequencing Topic model MIDI Megabyte Citizen science CycL

Software-defined networking Analytics GLONASS Dystopia

Multi-task learning Extensibility Blocks

5G RNA-Seq 3D printing Duplex TaqMan Rubric Python Smart grid AdaBoost

.bss Public records 6LoWPAN Internet of Things Pyrosequencing SPARQL

Crowdsourcing OpenFlow Memristor Green computing Optogenetics Radio

Omics Exome F1 score Small cell CRISPR Arduino Pre-clinical development

Machine to machine Lifelog Climateprediction.net Social media Net income

Multi-user MIMO Cyber-physical system Hardware security module

Cloud computing security Selective laser melting Malware Interquartile range

Broadcast communication network Visible light communication Android

Latent Dirichlet allocation Information sensitivity Rescue robot Display resolution

HTML5

Systems Modeling Language Three-dimensional integrated circuit Harmony search

Static single assignment form Cellulosic ethanol Visual odometry Nanopore Data set

Cascading failure Dashboard Geolocation

Artificial bee colony algorithm Agile software development Transmitter power output

Tiger Quadrature mirror filter

Homomorphic encryption Near field communication Bioconjugation

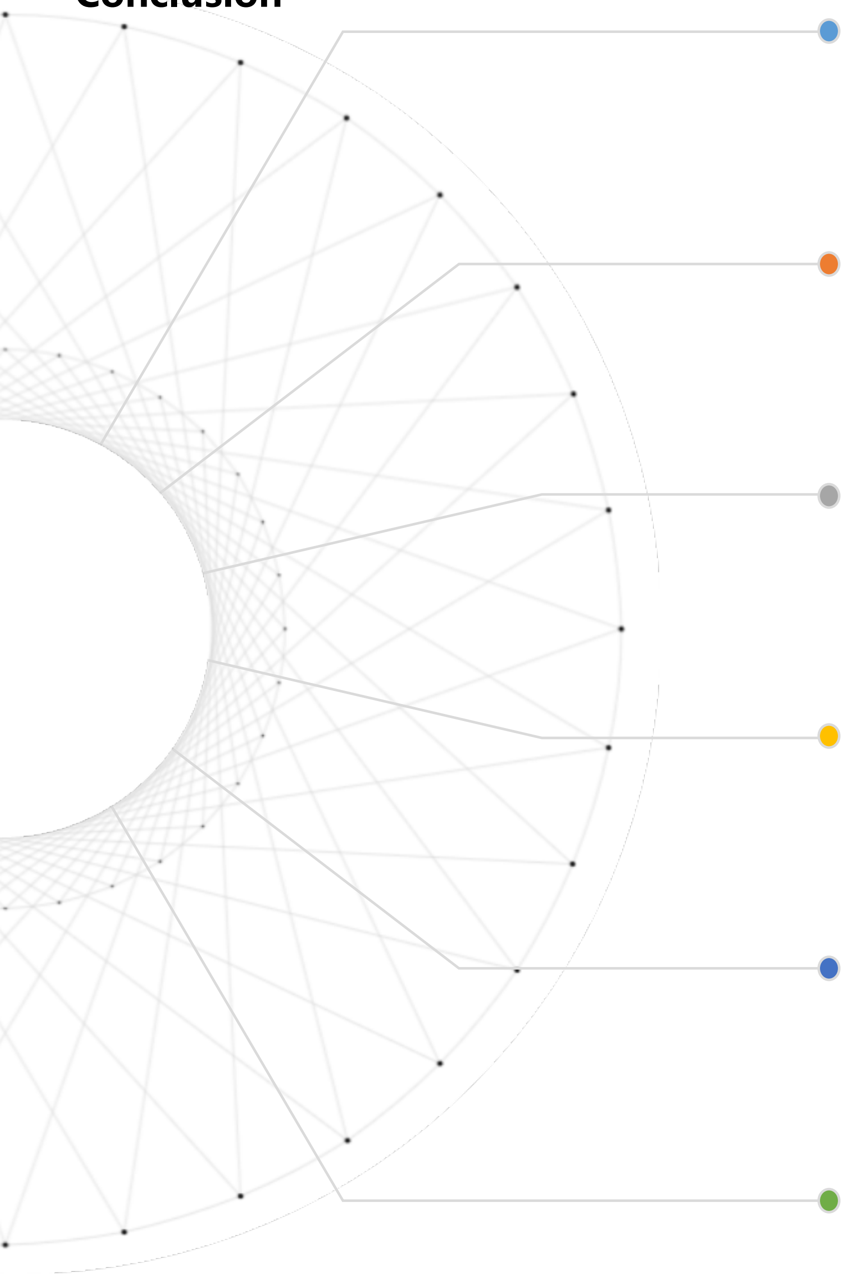
Microsoft Office Circulating tumor cell

Command-line interface Vehicular communication systems

Conclusion

PART FIVE

Conclusion



1 Insight the topic structure of CS field.

2 Find the connection between hot topics.

3 Extract the factors that can influence topics’ development.

4 Generate a time series dataset containing all 12243 topics

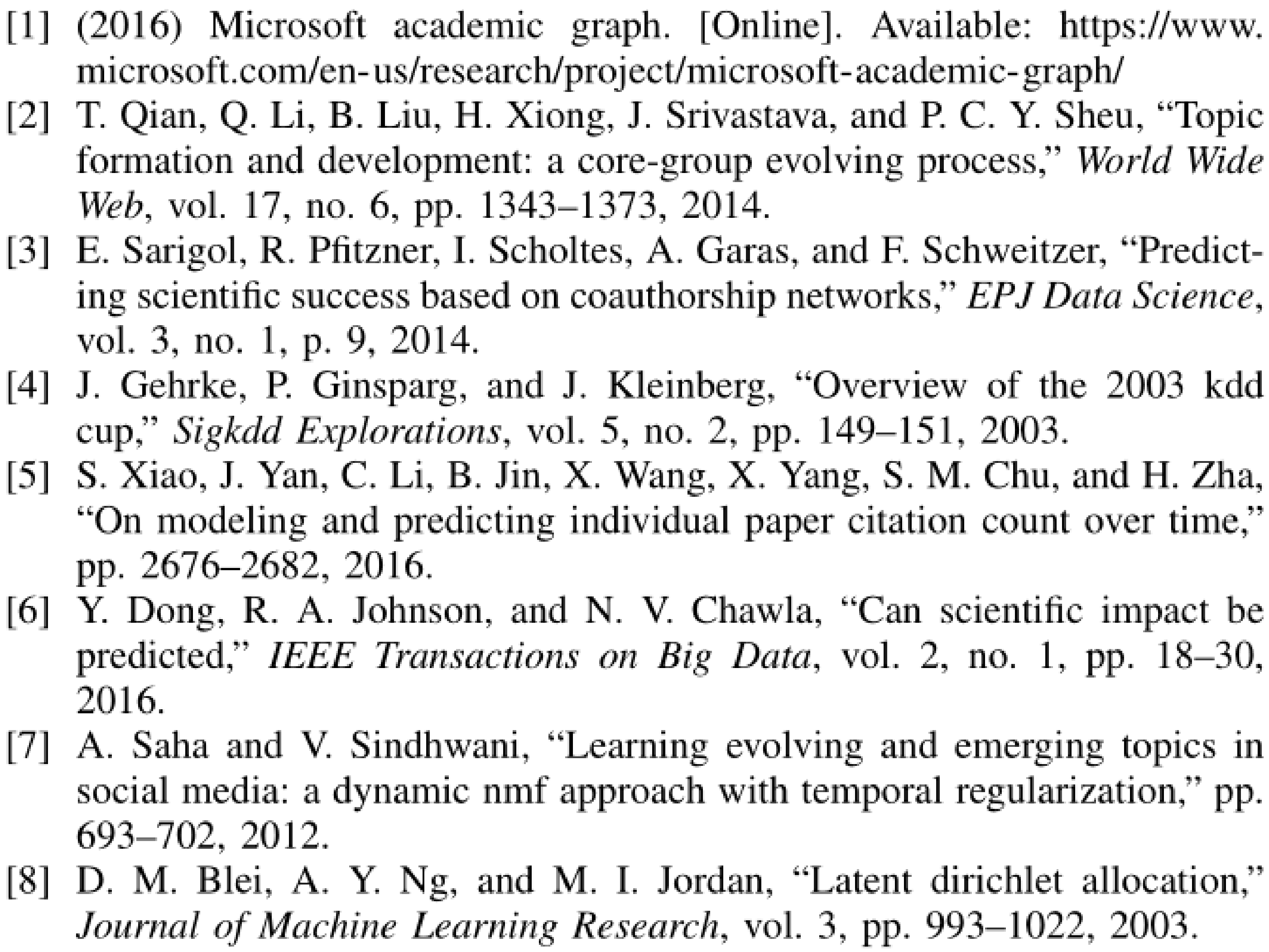
5 Compare the performance between different models.

6 Predict the top 100 hot topics in 5 years.

References

PART SIX

References



[1] (2016) Microsoft academic graph. [Online]. Available: https://www.microsoft.com/en-us/research/project/microsoft-academic-graph/

[2] T. Qian, Q. Li, B. Liu, H. Xiong, J. Srivastava, and P. C. Y. Sheu, "Topic formation and development: a core-group evolving process," *World Wide Web*, vol. 17, no. 6, pp. 1343–1373, 2014.

[3] E. Sarigol, R. Pfitzner, I. Scholtes, A. Garas, and F. Schweitzer, "Predicting scientific success based on coauthorship networks," *EPJ Data Science*, vol. 3, no. 1, p. 9, 2014.

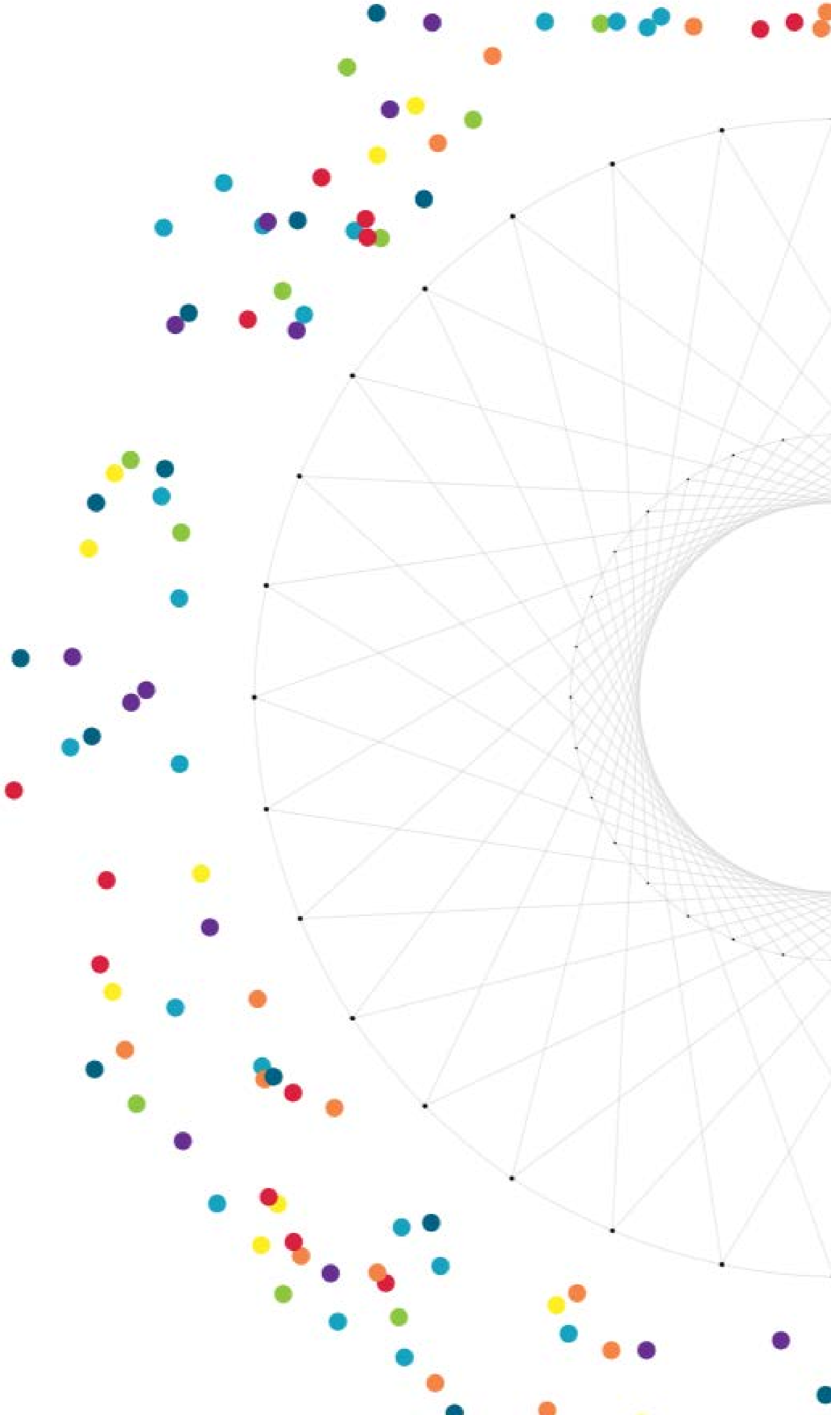
[4] J. Gehrke, P. Ginsparg, and J. Kleinberg, "Overview of the 2003 kdd cup," *Sigkdd Explorations*, vol. 5, no. 2, pp. 149–151, 2003.

[5] S. Xiao, J. Yan, C. Li, B. Jin, X. Wang, X. Yang, S. M. Chu, and H. Zha, "On modeling and predicting individual paper citation count over time," pp. 2676–2682, 2016.

[6] Y. Dong, R. A. Johnson, and N. V. Chawla, "Can scientific impact be predicted," *IEEE Transactions on Big Data*, vol. 2, no. 1, pp. 18–30, 2016.

[7] A. Saha and V. Sindhwani, "Learning evolving and emerging topics in social media: a dynamic nmf approach with temporal regularization," pp. 693–702, 2012.

[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.



Q&A

Thanks ！