Knowledge Graph: Wikipedia Computer Science

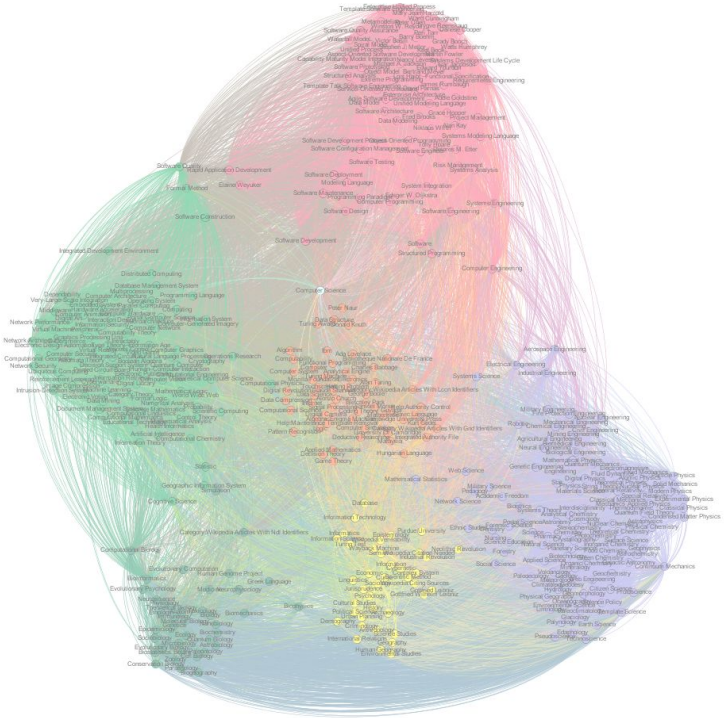
Christian Anyanwu, Trevor Larsen, Zack Mekus

Our Goal:

Create and analyze a knowledge graph of Computer Science using Wikipedia

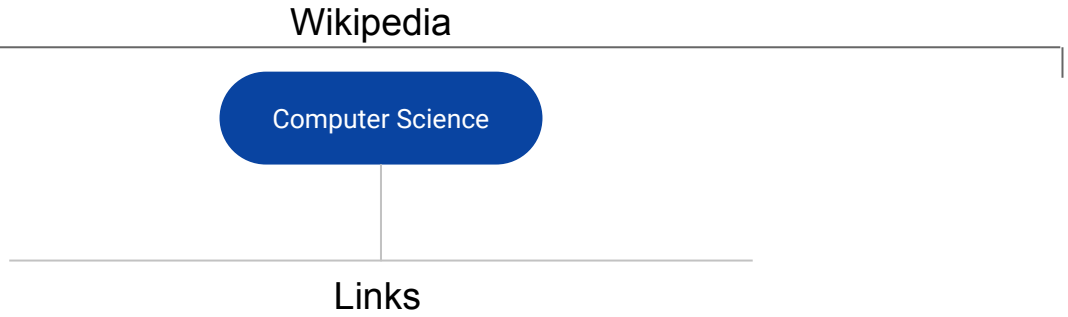
Project Overview

- Creation of graph
- Base Analysis
  - Node counts
  - Modularity Analysis
- Network Properties
  - In-degree
  - Out-degree
  - Closeness and Betweenness Centrality
  - Pagerank
- Machine Learning- Link Prediction



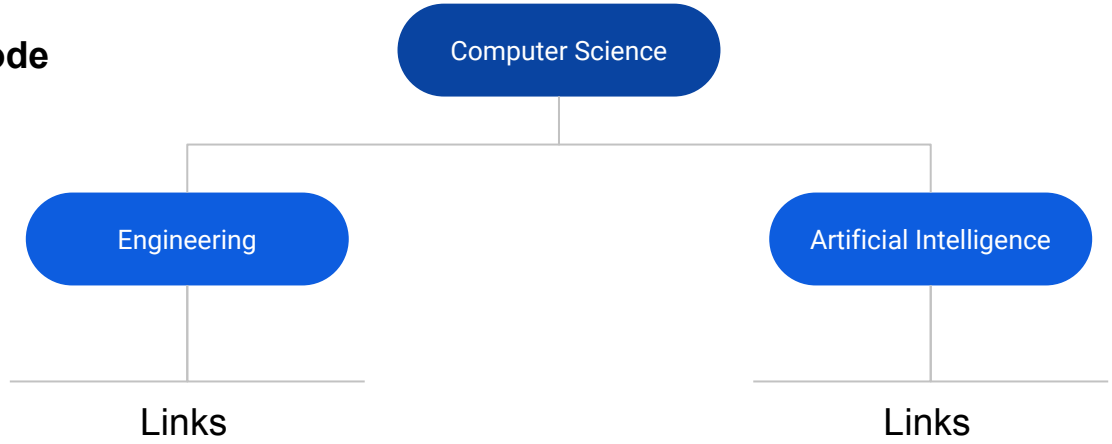
Creation of the Graph

Layer 1: Origin Node



Creation of the Graph

Layer 1: Origin Node



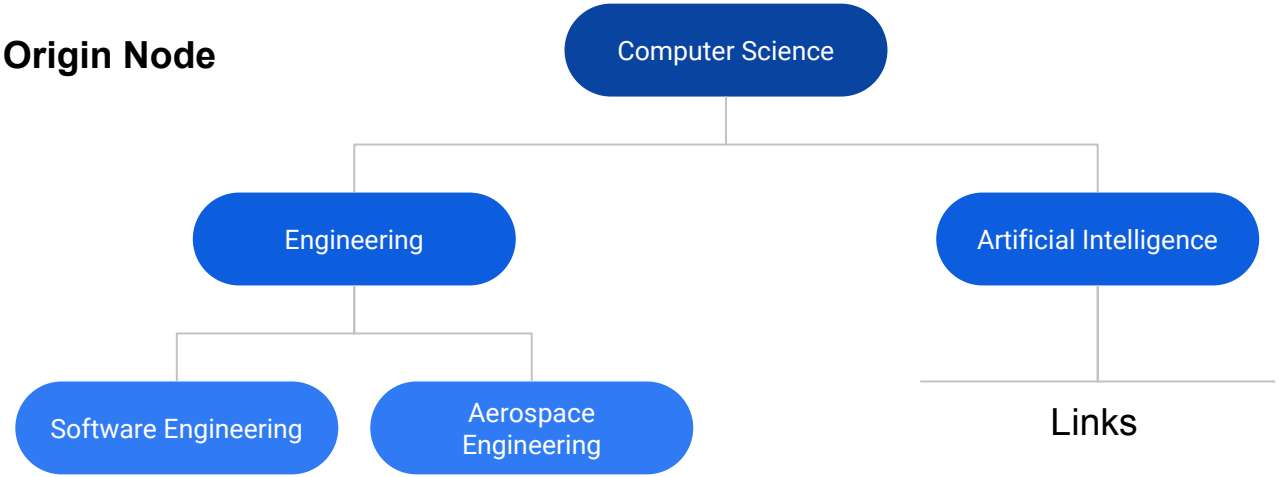
Layer 2:

Creation of the Graph

Layer 1: Origin Node

Layer 2:

Layer 3:

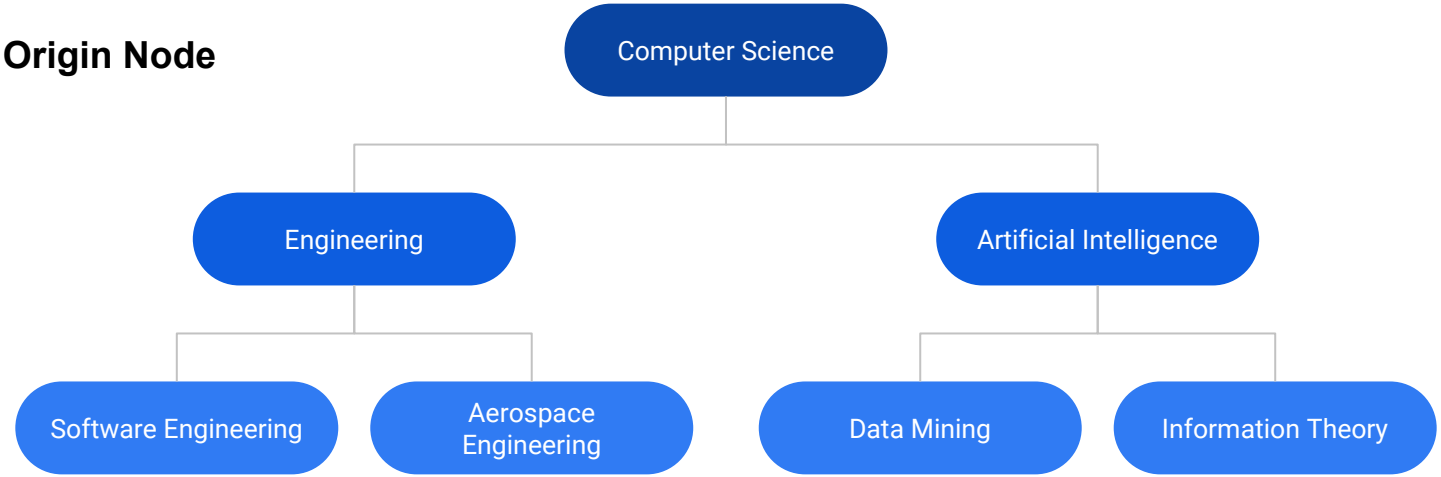


Creation of the Graph

Layer 1: Origin Node

Layer 2:

Layer 3:

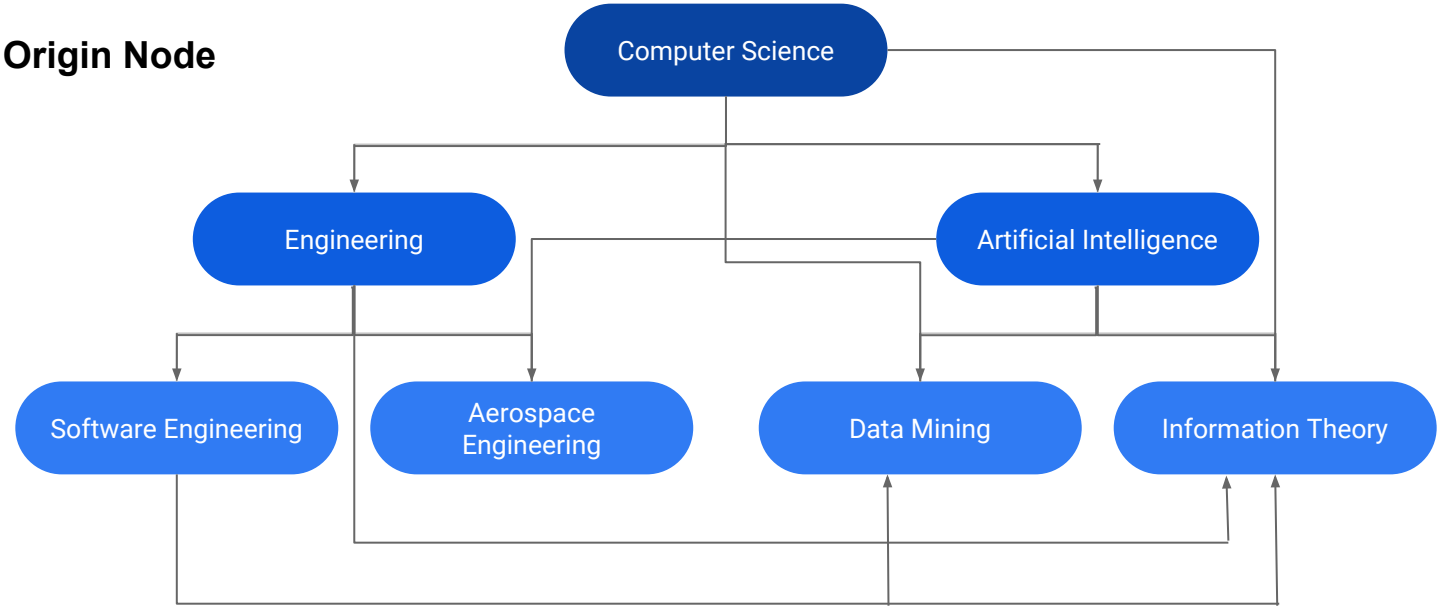


Creation of the Graph

Layer 1: Origin Node

Layer 2:

Layer 3:



What we found:

With just 3 layers:

# of Nodes = 22,099

# of Edges = 114,613

Avg. Degree = 10.36

What we found:

With just 3 layers:

# of Nodes = 22,099

# of Edges = 114,613

Avg. Degree = 10.36

Filter on Indegree (> 100)

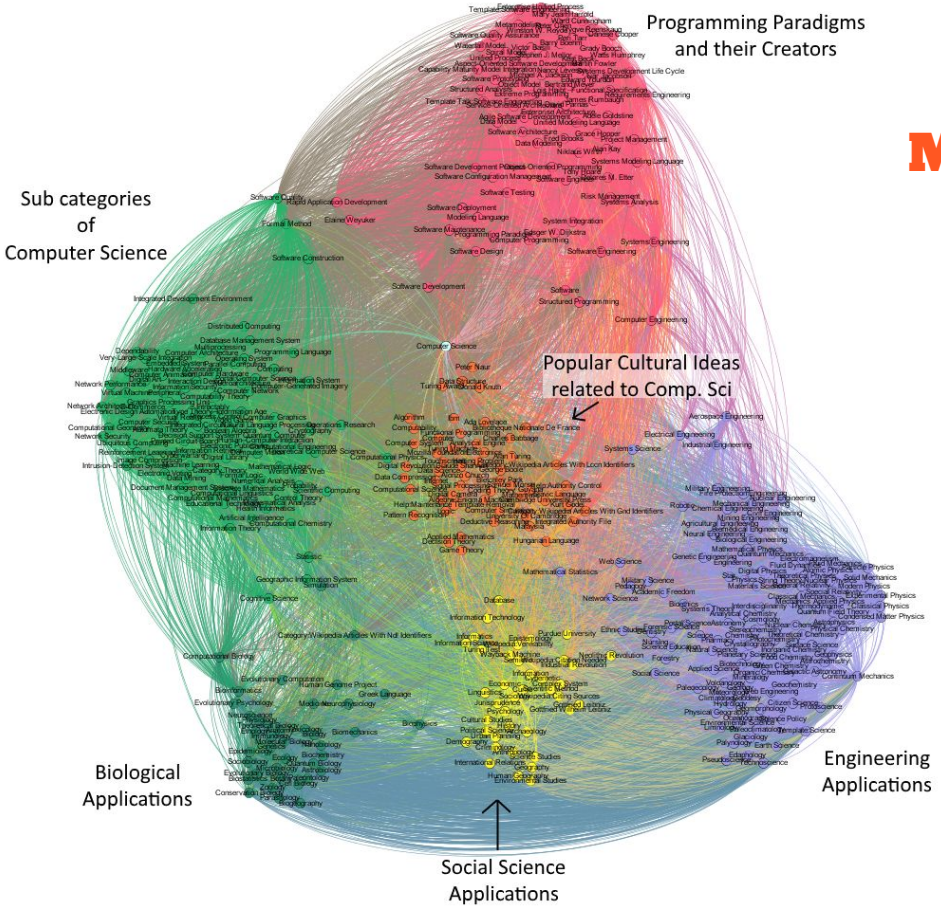
Final Graph:

# of Nodes = 418

# of Edges = 21,962

Avg. Degree = 105.53

Modularity Analysis



Main Takeaways:

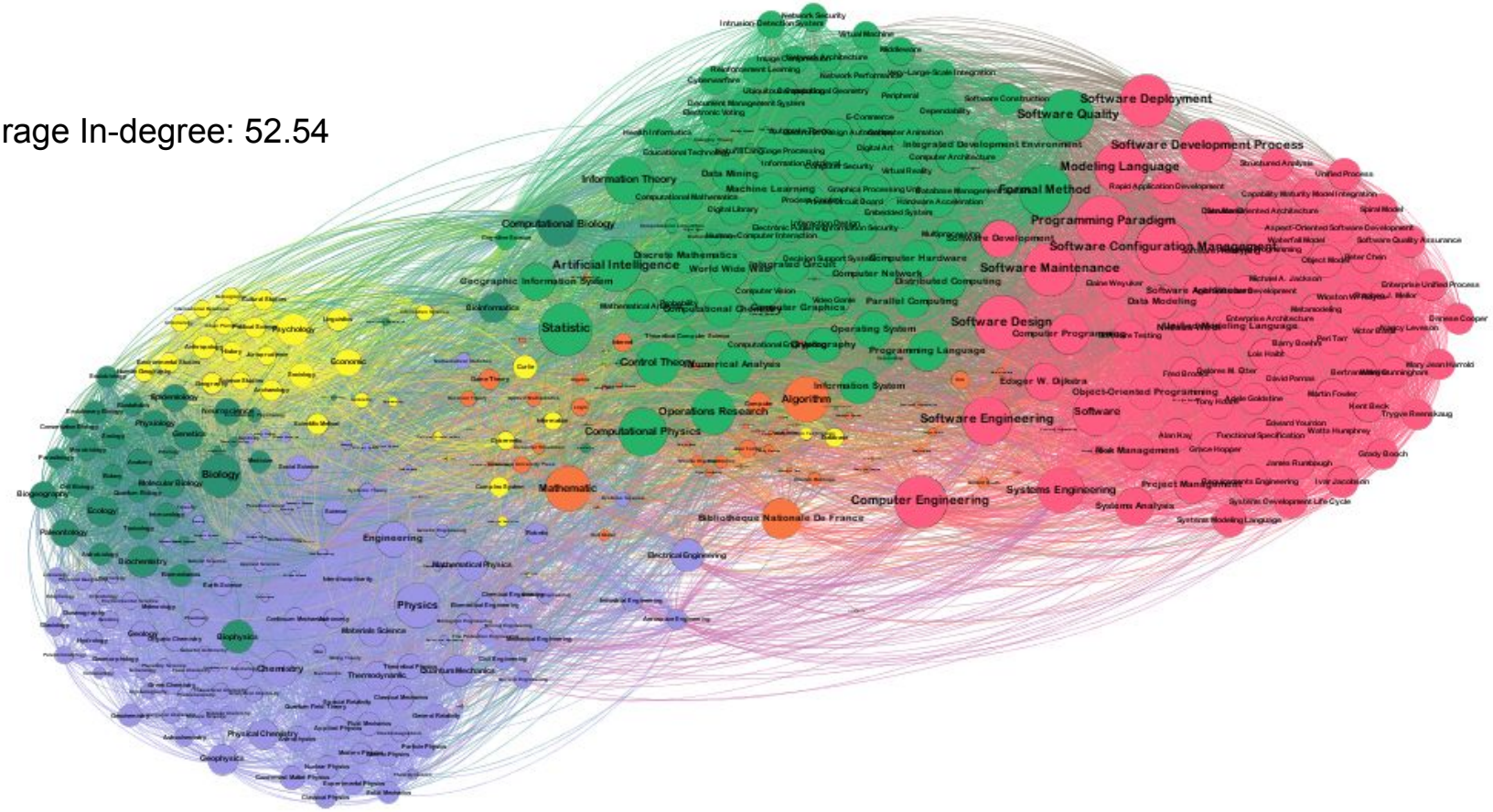
- Modularity Score: .457
- 7 Distinct Groups:
  1. Computer Science (Single Node)
  2. Subcategories of CS
  3. Programming Paradigms and their Creators
  4. Popular Cultural Ideas
  5. Social Science Applications
  6. Biological Applications
  7. Engineering Applications

Group Differences

	% of total Nodes	Avg. Degree	Avg. Clustering Coefficient
Engineering	27.16%	30.186	0.619
Subcategories	21.88%	41.549	0.588
Programming Paradigms	18.75%	63.615	0.861
Popular Cultural Ideas	13.22%	8.182	0.276
Social Science	9.38%	13.89	0.519
Biological	9.38%	21.46	0.717
Computer Science	0.24%	611	N/A

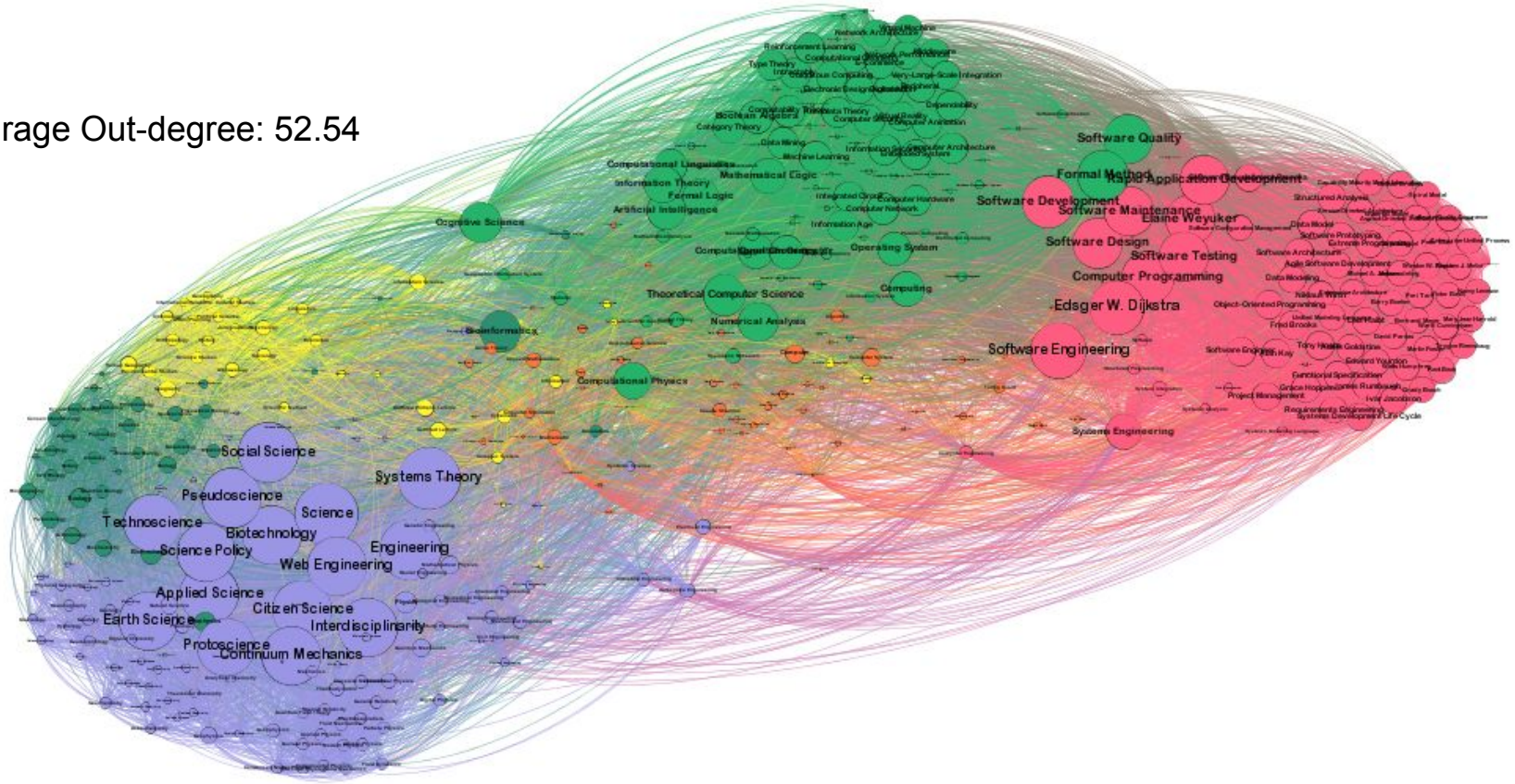
In-degree

Average In-degree: 52.54



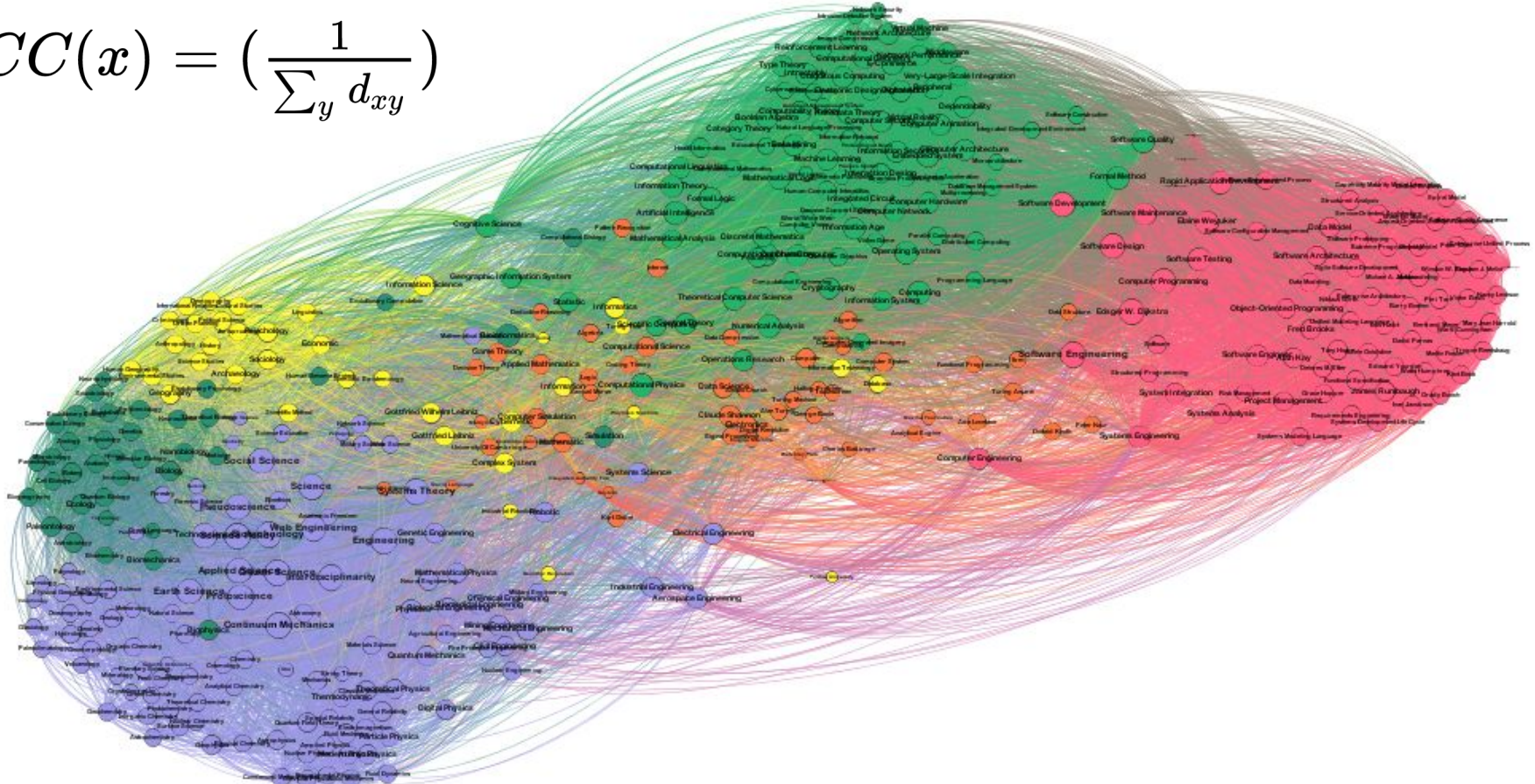
Out-degree

Average Out-degree: 52.54



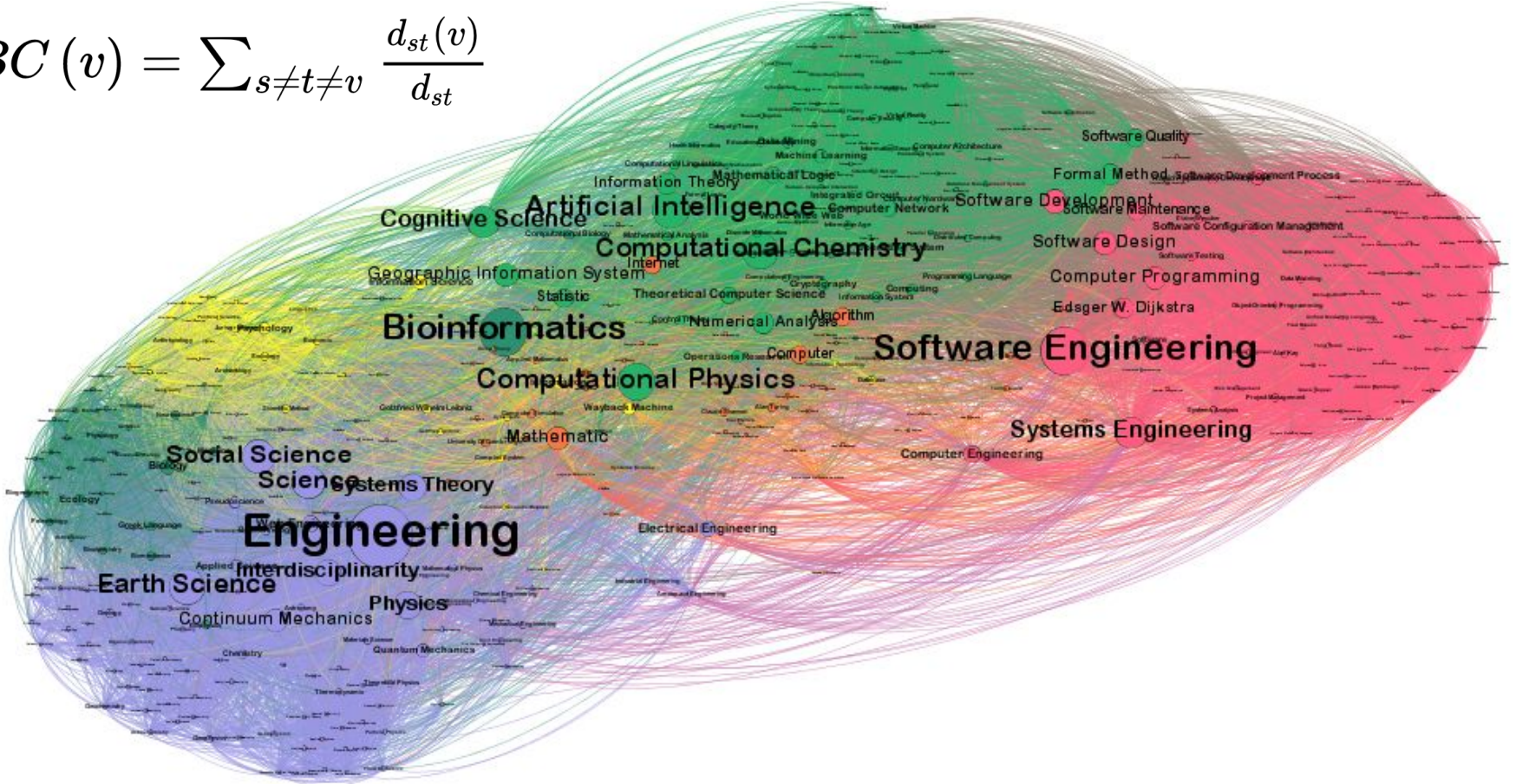
Closeness Centrality

$$CC(x) = (\frac{1}{\sum_y d_{xy}})$$

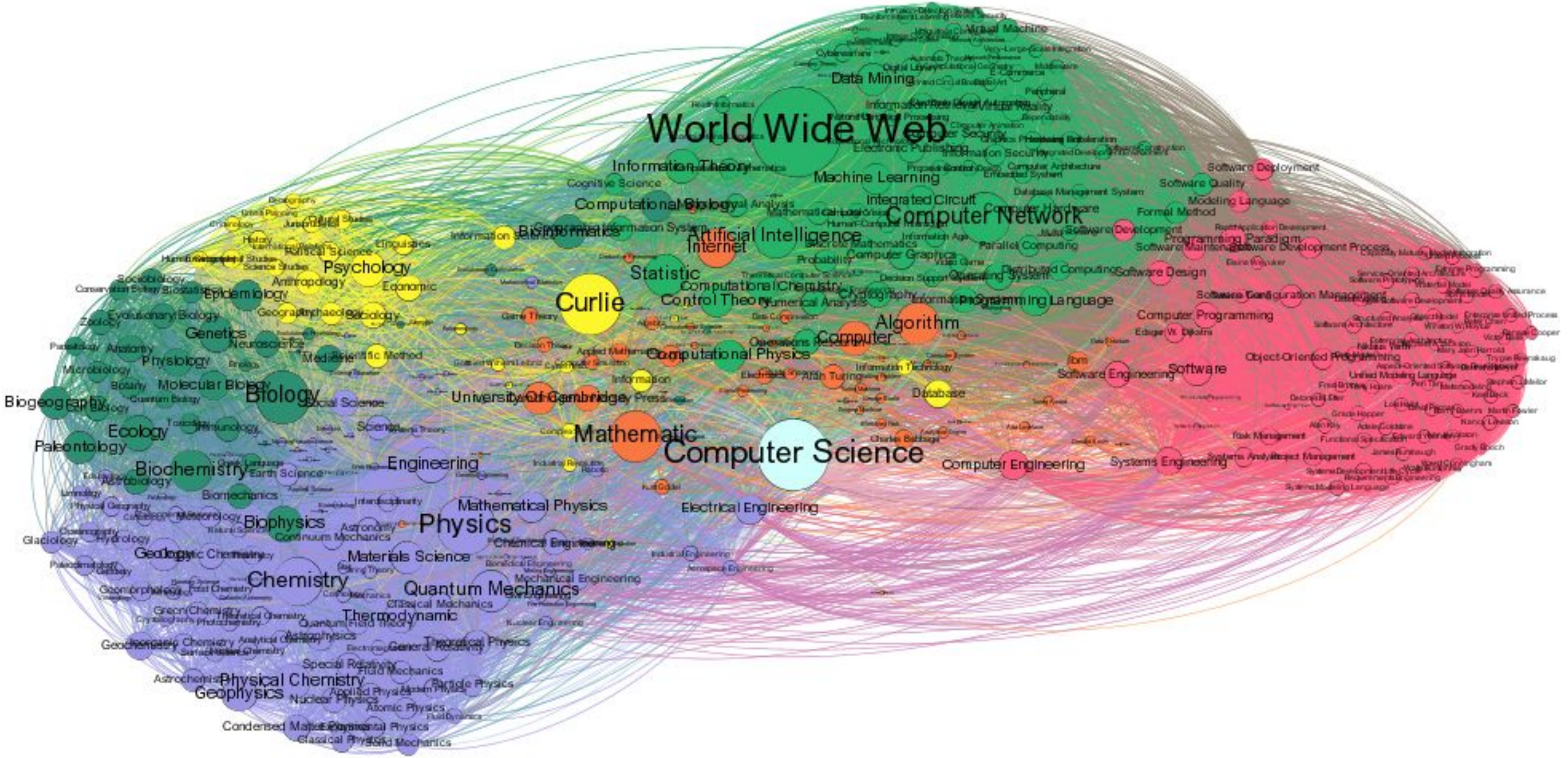


Betweenness Centrality

$$BC(v) = \sum_{s \neq t \neq v} \frac{d_{st}(v)}{d_{st}}$$



PageRank


World Wide Web
Computer Science
Curlie
Mathematic
Computer Network
Artificial Intelligence
Internet
Statistic
Algorithm
Computer
Physics
Chemistry
Biology
Biochemistry
Quantum Mechanics
Physical Chemistry
Geophysics
Psychology
Epidemiology
Genetics
Ecology
Paleontology
Biogeography
Biophysics
Engineering
Mathematical Physics
Materials Science
Thermodynamic
Electrical Engineering
Computer Engineering
Systems Engineering
Software Engineering
Software
Object-Oriented Programming
Computer Programming
Programming Language
Machine Learning
Data Mining
Information Theory
Computational Physics
Computational Chemistry
Control Theory
Database
Software Deployment
Software Quality
Modeling Language
Formal Method
Programming Paradigm
Software Design
Parallel Computing
Computer Graphics
Integrated Circuit
Cognitive Science
Bioinformatics
Anthropology
Economic
Linguistics
Physiology
Anatomy
Molecular Biology
Microbiology
Zoology
Evolutionary Biology
Neuroscience
Medicine
Scientific Method
Science
Earth Science
Biomechanics
Astronomy
Continuum Mechanics
Glaciology
Hydrology
Geology
Geomorphology
Classical Mechanics
Quantum Field Theory
Theoretical Physics
Special Relativity
Fluid Mechanics
Particle Physics
Nuclear Physics
Atomic Physics
Condensed Matter Physics
Solid Mechanics
Mechanical Engineering
Electronic Publishing
Virtual Machine
Probability
Numerical Analysis
Information Technology
Alan Turing

Link Prediction

Predict if a wikipedia page will contain a link to another wikipedia page

163,620 possible links

12.6% actual links

Use Logistic Regression

The Plan

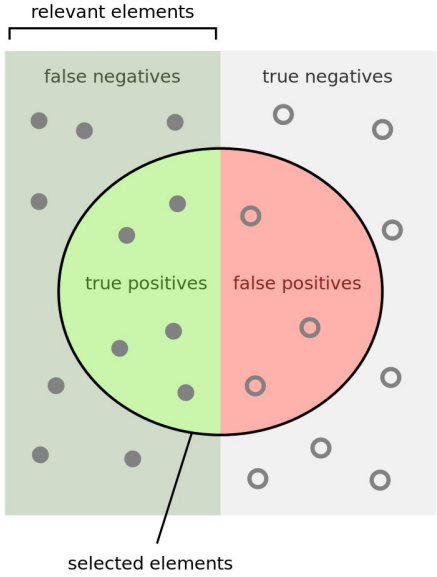
Define feature set between each pair of nodes

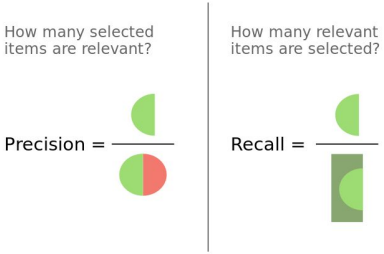
Use Logistic Regression

Train on 90% of the data and test on 10%

Use F1 score to evaluate

$$F1 = \left(\frac{precision^{-1} + recall^{-1}}{2}\right)^{-1}$$





https://en.wikipedia.org/wiki/F1_score

Initial Feature Set

Outdegree of source node

Indegree of target node

Example

Does Network Science have a link to Data Mining?

Network Science outdegree: 7

Data Mining indegree: 71

Calculation

	Bias	Source Outdegree	Target Indegree
Value	1	7	71
Weight	-3.99	.018	.014
Contribution	-3.99	0.13	0.99

Sum of contributions = -2.93

Probability of Link = 5%

Link Prediction = No

Initial Results

	Actually NOT Linked	Actually Linked
Predicted NOT Linked	14031	1724
Predicted Linked	348	259

Precision: 0.13

$$precision = (\frac{true\ positive}{classified\ positive})$$

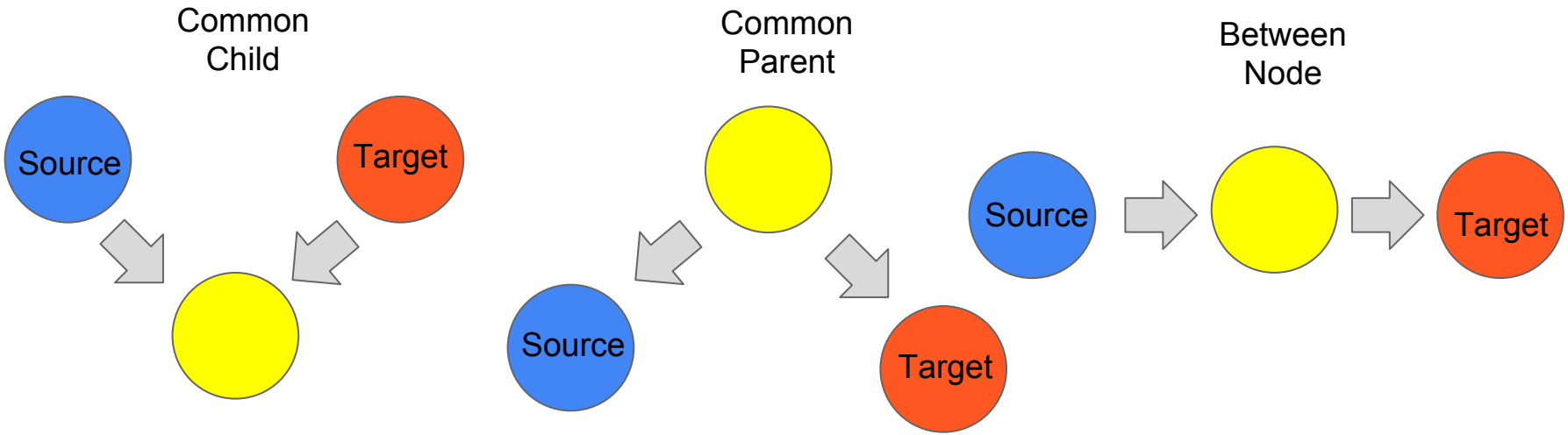
Recall: 0.43

$$recall = (\frac{true\ positive}{actually\ positive})$$

F1 Score: 0.20

$$F1 = (\frac{precision^{-1} + recall^{-1}}{2})^{-1}$$

Local Graph Measures



Example

Network Science and Data Mining

Both Link to:	Computer Networks, Statistics
Both are linked to from:	Computer Networks, Computer Science
Between the two:	Computer Networks, Statistics

Calculation

	Bias	Outdegree	Indegree	Children	Parents	Between
Value	1	7	71	2	2	2
Weight	-4.12	-.007	-.011	.017	.025	.24
Contribution	-4.12	-0.05	-0.78	0.03	0.05	0.50

Sum of contributions = -4.37

Probability of Link = 1%

Link Prediction = No

Results

	Actually NOT Linked	Actually Linked
Predicted NOT Linked	14274	498
Predicted Linked	105	1485

Precision: 0.93

$$precision = (\frac{true\ positive}{classified\ positive})$$

Recall: 0.75

$$recall = (\frac{true\ positive}{actually\ positive})$$

F1 Score: 0.83

$$F1 = (\frac{precision^{-1} + recall^{-1}}{2})^{-1}$$

Domain Specific Measures

Categories that are common to both wikipedia articles

Words in the first 300 characters that are common to both articles and aren’t stopwords

Example

Network Science and Data Mining

Common Categories: “All articles with unsourced statements”

Common Words: computer, science

Calculation

	Bias	Outdegree	Indegree	Children	Parents	Between	Category	Words
Value	1	7	71	2	2	2	1	2
Weight	-4.87	-.007	-.013	.025	.036	.20	.20	.43
Contribution	-4.87	-.05	-.92	.05	.07	.40	.20	.86

Sum of contributions = -4.26

Probability of Link = 2%

Link Prediction = No

Results

	Actually NOT Linked	Actually Linked
Predicted NOT Linked	14273	464
Predicted Linked	106	1519

Precision: 0.93

$$precision = (\frac{true\ positive}{classified\ positive})$$

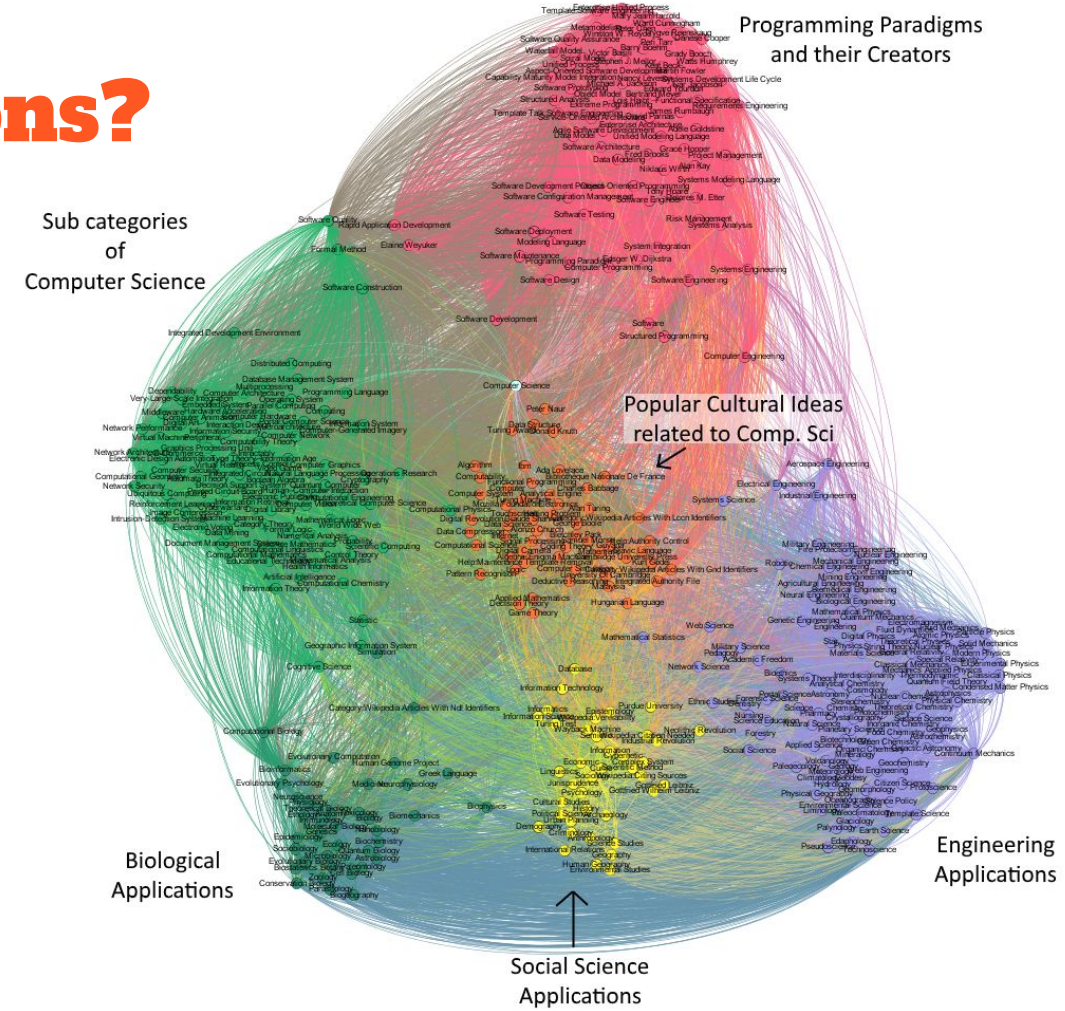
Recall: 0.77

$$recall = (\frac{true\ positive}{actually\ positive})$$

F1 Score: 0.84

$$F1 = (\frac{precision^{-1} + recall^{-1}}{2})^{-1}$$

Questions?


Programming Paradigms and their Creators
Sub categories of Computer Science
Popular Cultural Ideas related to Comp. Sci
Biological Applications
Engineering Applications
Social Science Applications