

CS247: ADVANCED DATA MINING

1: Introduction

Instructor: Yizhou Sun

yzsun@cs.ucla.edu

January 3, 2022

Course Information

- Course homepage:
https://github.com/yichousun/Winter2022_CS247_AdvDM
- Class Schedule
 - Slides
 - Assignments
 - Course projects
 - ...

• Prerequisites

- Required prerequisite courses: CS 145 or CS 146 or equivalent.
- You are expected to have background knowledge in data structures, algorithms, basic linear algebra, and basic statistics.
- You are expected to know basic knowledge in data mining and machine learning.
- You will also need to be familiar with at least one programming language, and have programming experiences.

Meeting Time and Location

- When
 - M&W, 2-3:50pm
- Where
 - Zoom for the first two weeks:
<https://ucla.zoom.us/j/94865976682?pwd=MTZqM054aG1NeXVhVIFmbnV1WExNQQT09>
 - Meeting ID: 948 6597 6682
 - Passcode: 447984
 - Classroom
 - BOELTER 4760

Instructor and TA Information

- Instructor: Yizhou Sun
 - Homepage: <http://web.cs.ucla.edu/~yzsun/>
 - Email: yzsun@cs.ucla.edu
 - Office: 3531F
 - Office hour: Mondays/Tuesdays 4-5pm @ same zoom as lecture
- TA: Zongyue Qin
 - Email: qinzongyue@cs.ucla.edu
 - Office hours: Tuesday 12-13:00 and Thursday 11:00-12:00 (Zoom can be found at canvas)

Discussion Session

- Friday 2-3:50pm
- Classroom: Kinsey Science Teaching Pavilion
1220B

Grading

- Homework: 50%
- Quizzes: 15%
- Course project: 35%

Grading: Homework

- Homework: 50%
 - 6 assignments are expected (highest 5 scores will be picked)
 - Bonus questions might be included, which can be carried over to the final score
 - Deadline: 11:59pm of the indicated due date via *Bruinlearn* system
 - *Late submission policy*: get original score* if you are t hours late.
- **No copying or sharing of homework!**
 - But you can discuss general challenges and ideas with others
 - *Suspicious cases will be reported to The Office of the Dean of Students*

$$\mathbf{1}(t \leq 24)e^{-(\ln(2)/12)*t}$$

Grading: Quizzes

- Quizzes: 15%
 - Popup In-class quizzes
 - For Session LEC 1:
 - 24 hours extension to accommodate different time zones
 - For Session LEC 80:
 - A separate link with deadline on the following Sunday midnight (11:59pm)
- Get you in track 😊
 - No reminders

Grading: Course Project

- Course project: 35%
 - Group project (3-4 people for one group)
 - Goal: Solve an open data mining problem, related to text, graph/networks, or recommender systems
 - You are expected to present your project to the class, and submit a project report and your code at the end of the quarter
 - **Warning: private peer review will be collected in the end, and free rider will receive a significant lower score**

slido



**Join at slido.com
#146811**

ⓘ Start presenting to display the joining instructions on this slide.

slido



Audience Q&A Session

ⓘ Start presenting to display the audience questions on this slide.

Goals of the Course

- Review and understand fundamentals of basic data mining techniques
- Learn recent data mining techniques on several advanced data types
- Develop skills to apply data mining algorithms to solve real-world applications
- Gain initial experience in conducting research on data mining

Content to be Covered

- Part I: Basics of data mining
- Part II: Text mining
- Part III: Graphs/Network mining
- Part IV: Recommender systems

Part I: Basics of data mining (~2 weeks)

- Basics of Probabilistic Models
- K-means and mixture models
- Neural networks and deep learning

Part II: Text mining (~2 weeks)

- Topic Models
- Word Embedding

Text Data

- “Text mining, also referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. High-quality information is typically derived through the devising of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers to some combination of relevance, novelty, and interestingness. Typical text mining tasks include text categorization, text clustering, concept/entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling (i.e., learning relations between named entities).” –from wiki

Text Data – Topic Modeling

Topics

gene 0.04
dna 0.02
genetic 0.01
...

life 0.02
evolve 0.01
organism 0.01
...

brain 0.04
neuron 0.02
nerve 0.01
...

data 0.02
number 0.02
computer 0.01
...

Documents

Seeking Life's Bare (Genetic) Necessities

COLD SPRING HARBOR, NEW YORK— How many **genes** does an **organism** need to **survive**? Last week at the genome meeting here,* two genome researchers with radically different approaches presented complementary views of the basic genes needed for **life**. One research team, using **computer** analyses to compare known **genomes**, concluded that today's **organisms** can be sustained with just 250 genes, and that the earliest life forms required a mere 128 **genes**. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those **predictions**

"are not all that far apart," especially in comparison to the 75,000 **genes** in the human genome, notes Siv Andersson at Uppsala University in Sweden. "They arrived at the 800 number. But coming up with a consensus answer may be more than just a **genetic** numbers game, particularly if more and more **genomes** are completely mapped and sequenced. "It may be a way of organizing any newly **sequenced genome**," explains Arcady Mushegian, a **computational** molecular biologist at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing an

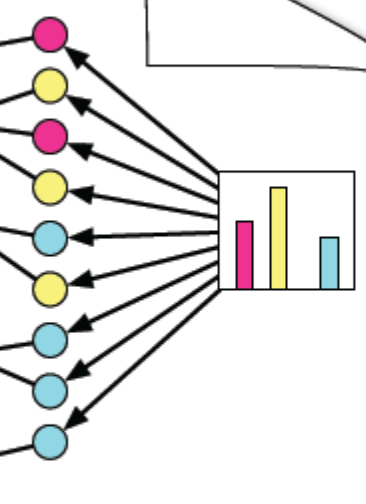


* Genome Mapping and Sequencing, Cold Spring Harbor, New York, May 8 to 12.

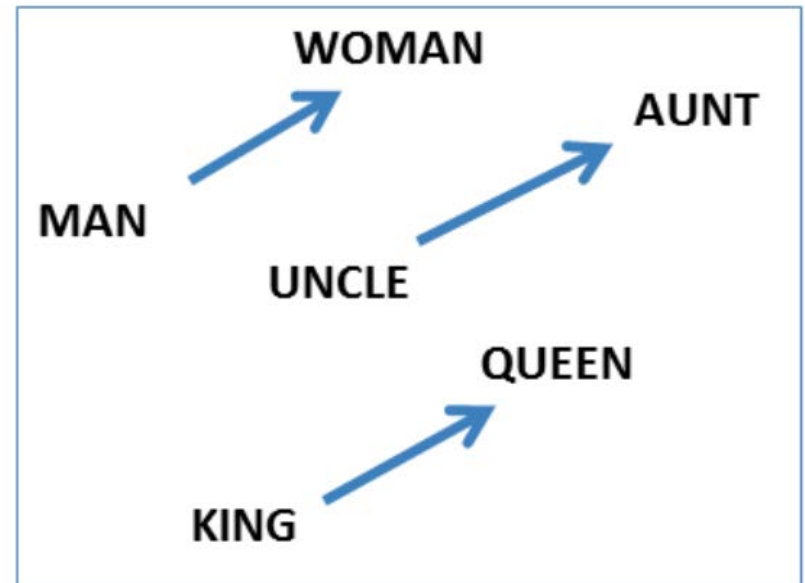
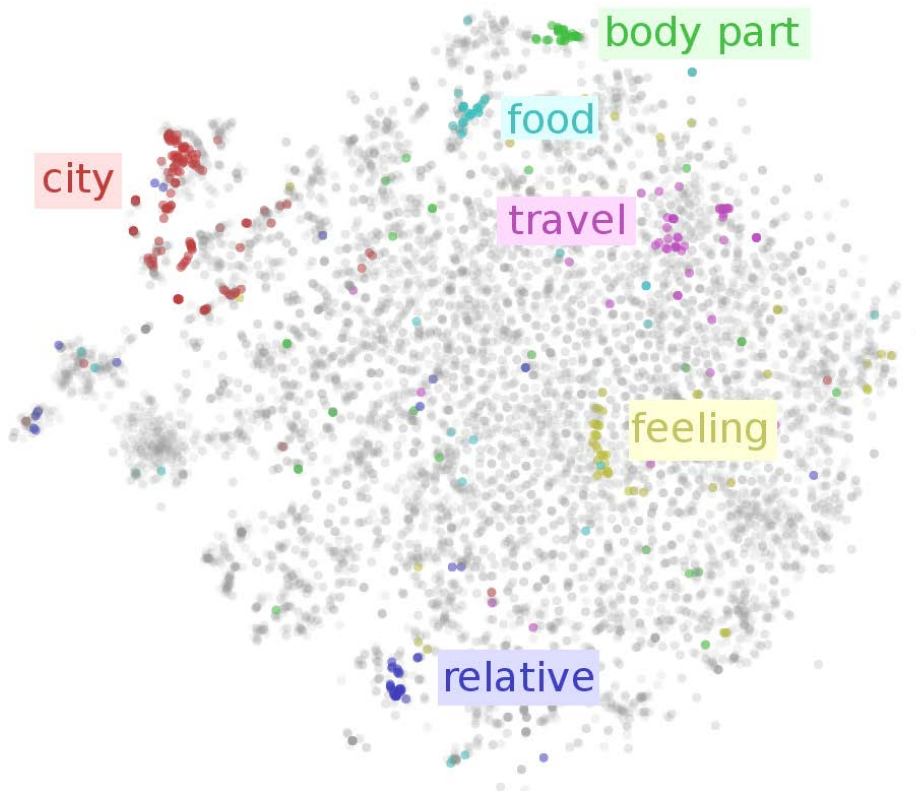
Stripping down. Computer analysis yields an estimate of the minimum modern and ancient genomes.

SCIENCE • VOL. 272 • 24 MAY 1996

Topic proportions and assignments



Text Data – Word Embedding

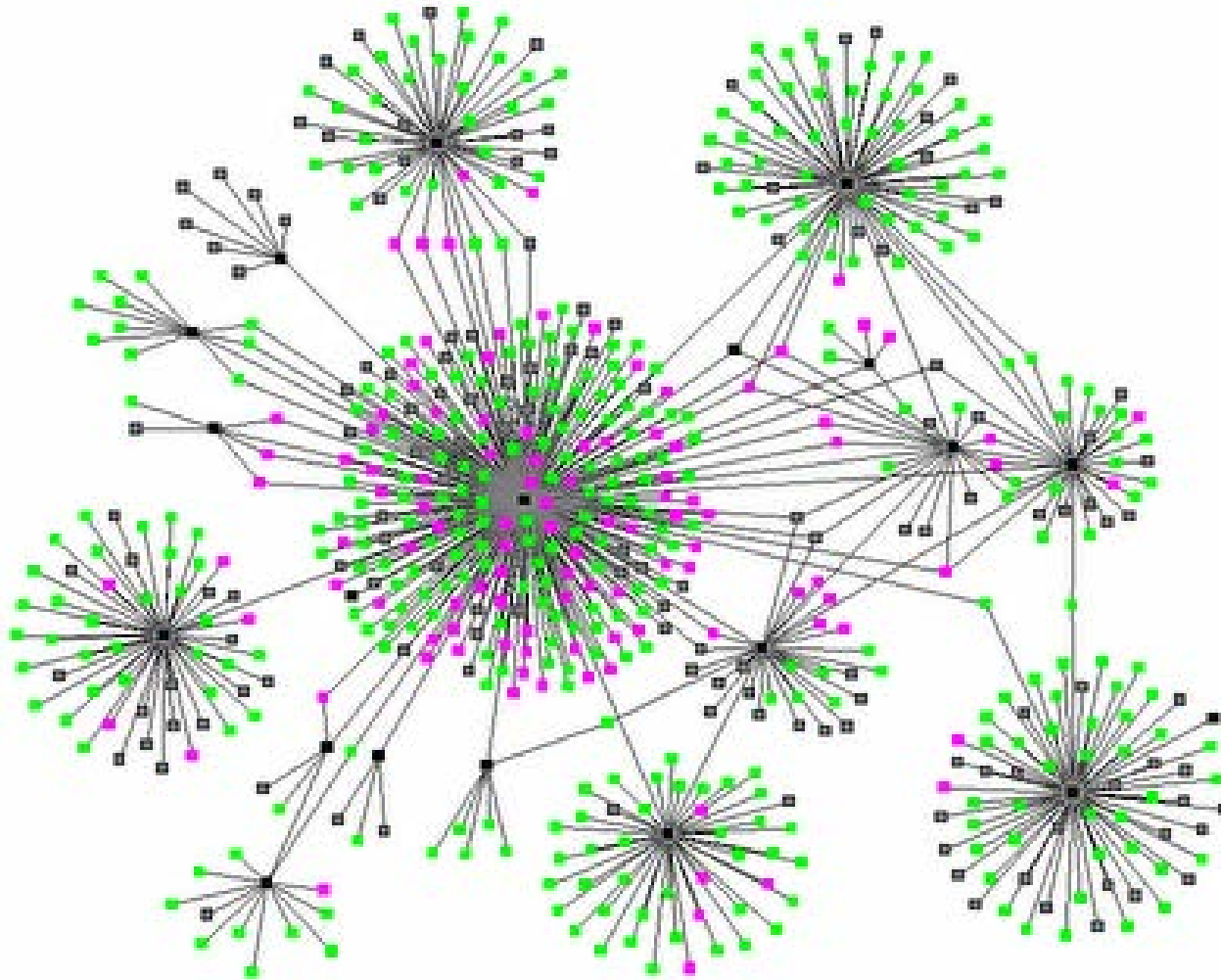


king - man + woman = queen

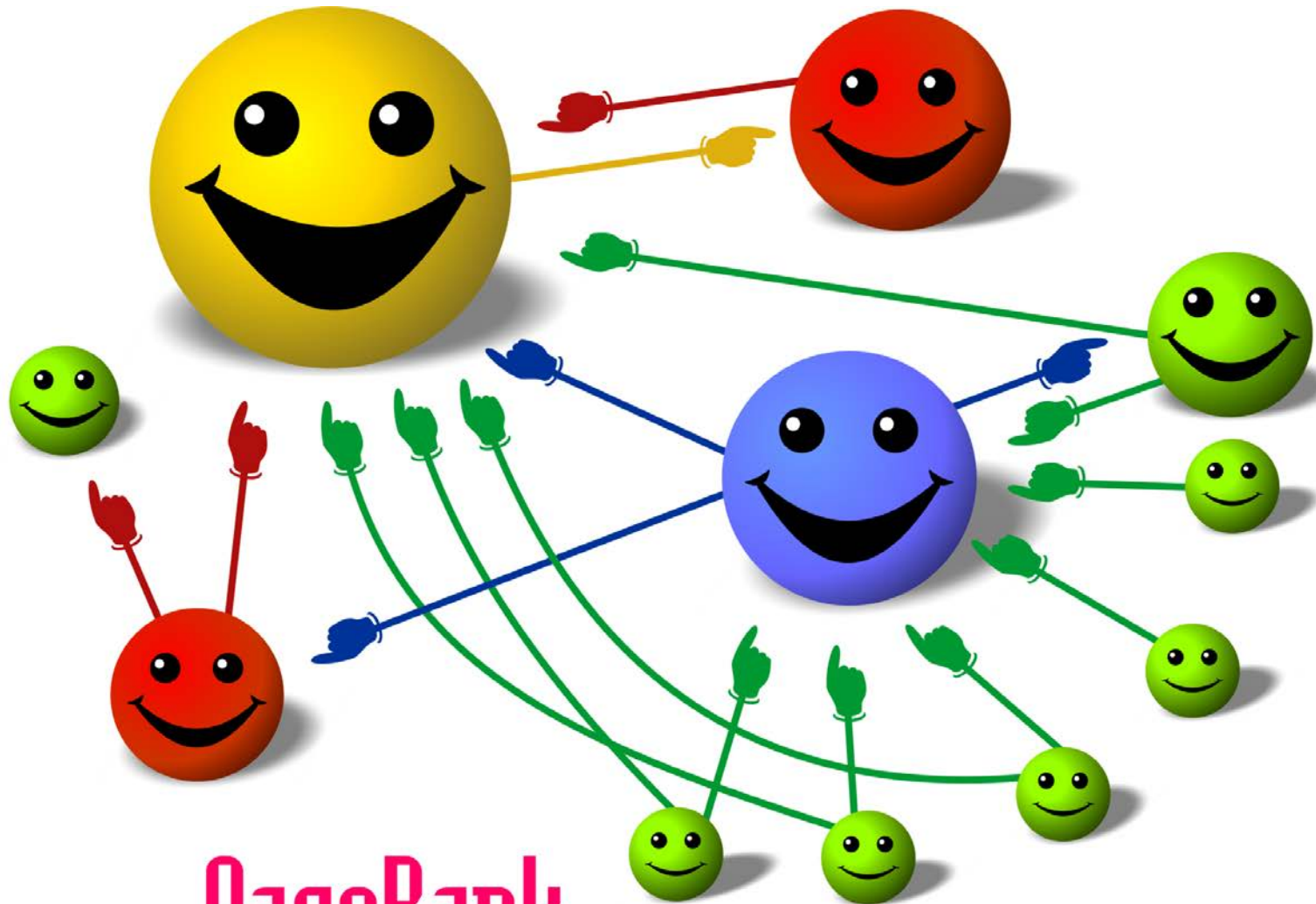
Part III: Graphs/Network mining (~2.5 weeks)

- Random walk on graphs
- Spectral clustering
- Label propagation
- Graph/Network embedding
- Knowledge graph embedding

Graph / Network



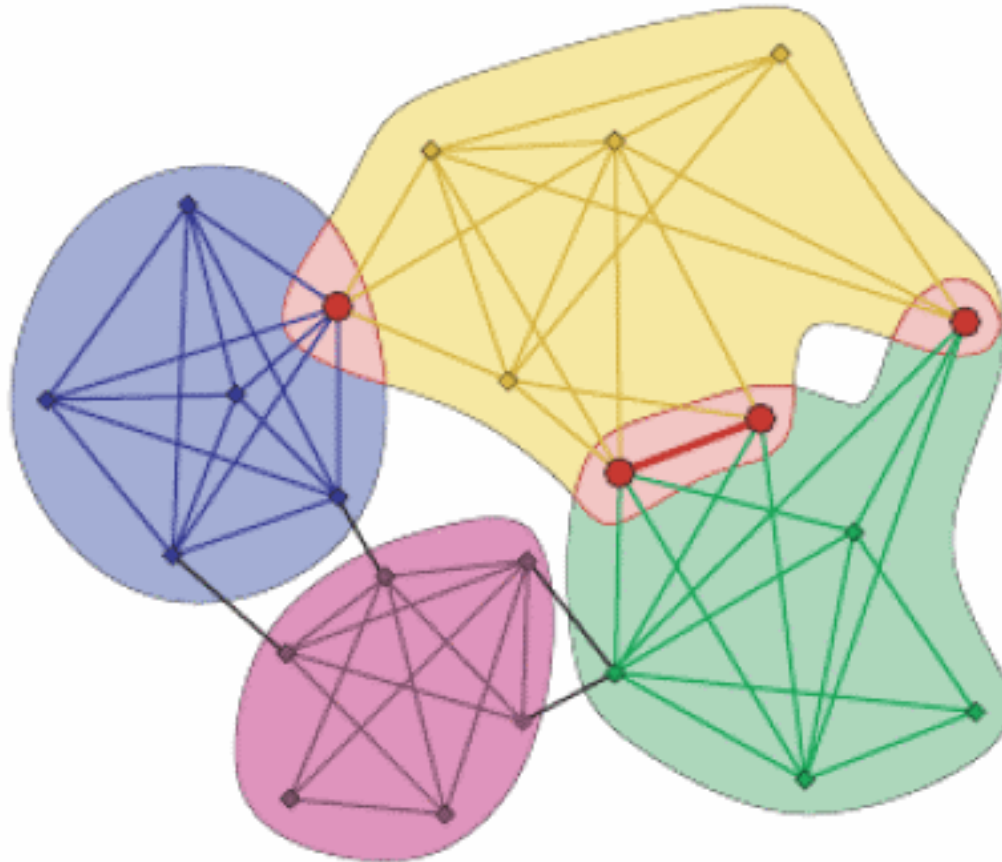
PageRank



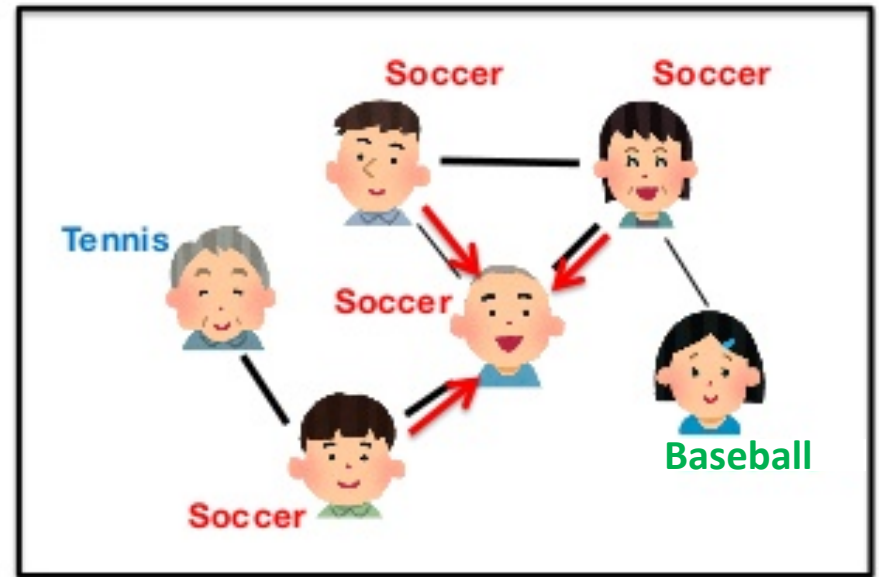
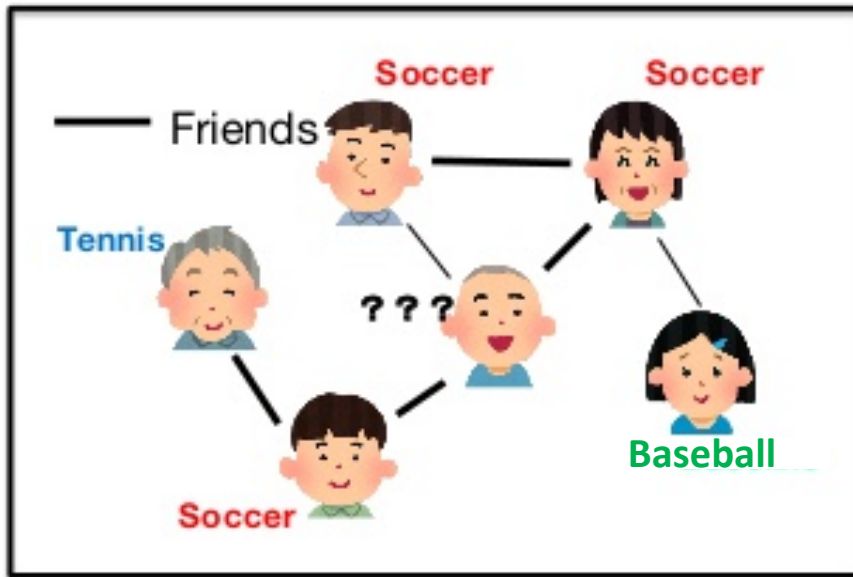
PageRank

Source: wikipedia

Community Detection

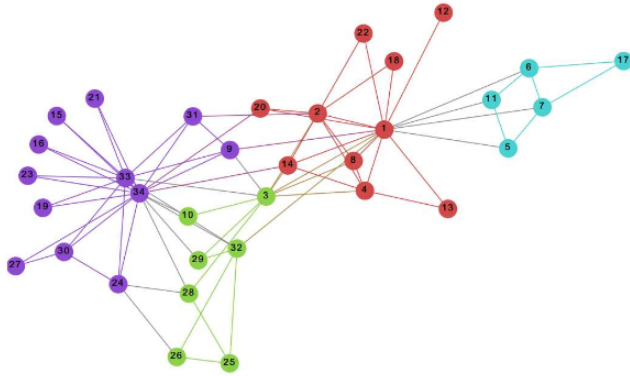


Label Propagation

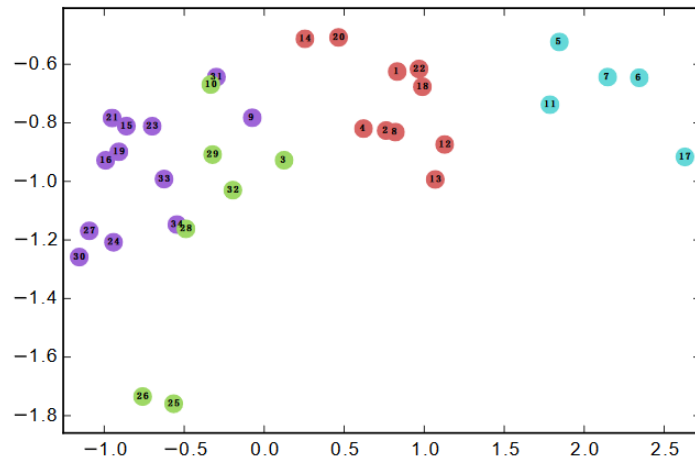


Source: <https://www.slideshare.net/yamaguchiyuto/when-does-label-propagation-fail-a-view-from-a-network-generative-model>

Network Embedding



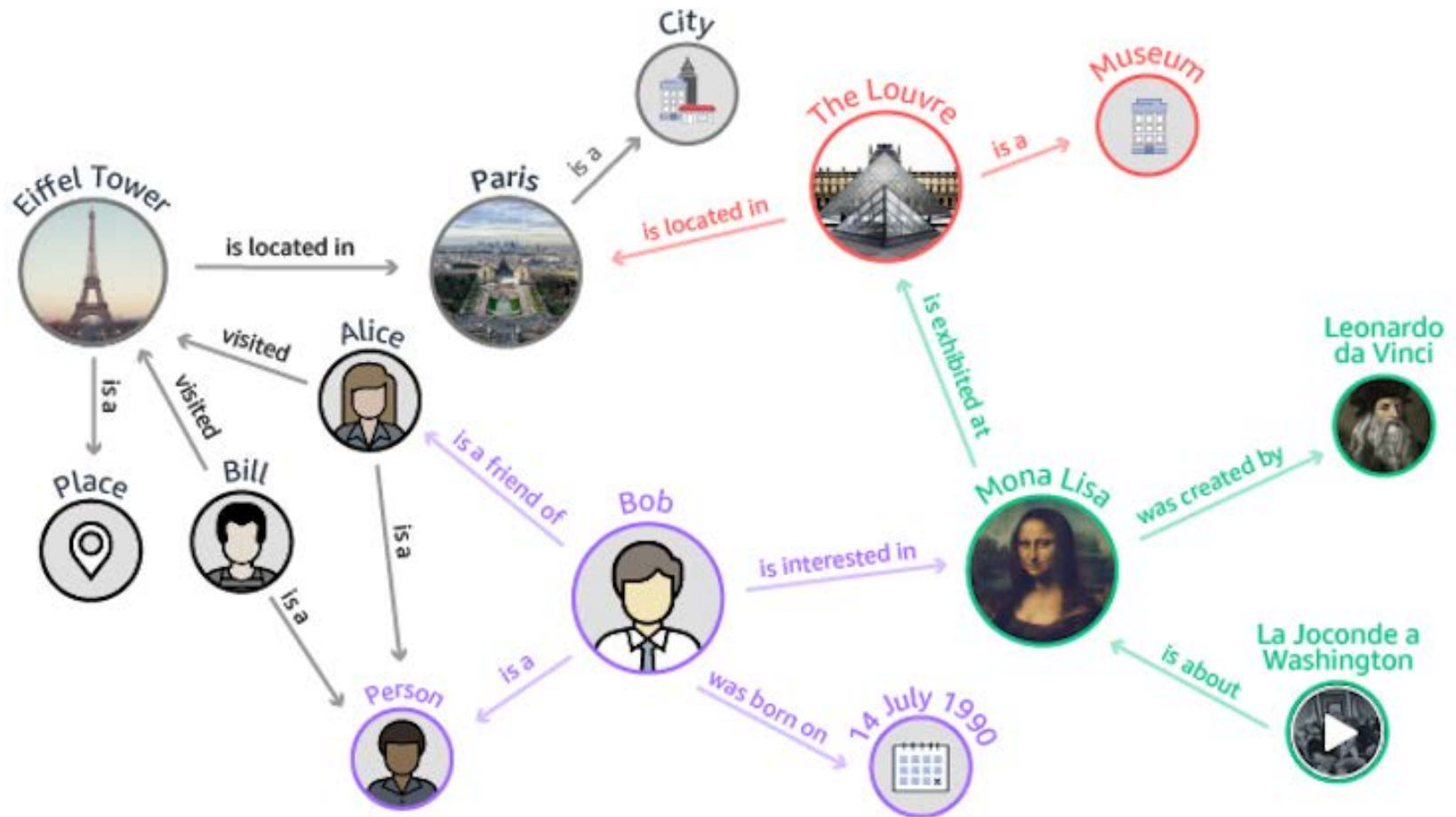
(a) Input: karate network



(b) Output: representations

Source: DeepWalk

Knowledge Graph



<https://geomarketing.com/amazons-neptune-database-will-expand-the-knowledge-graph-heres-how/amazon-neptune-knowledge-graph>

Part IV: Recommender systems (~2 weeks)

- Collaborative filtering, matrix factorization, Bayesian personalized ranking
- Factorization machine and deep collaborative filtering
- Recommendation from graph perspective

Recommender systems

You may also like



Jack & Jones
JAMIE - Polo shirt - orange
£21.00

Free delivery & returns

ALTERNATIVE PRODUCTS

Beko Washing Machine

Code: WMB81431LW

£269.99

Zanussi Washing Machine

Code: ZWH6130P

£269.99

Blomberg Washing Machine

Code: WNF6221

£299.99

You may also like



★★★★☆ (109)



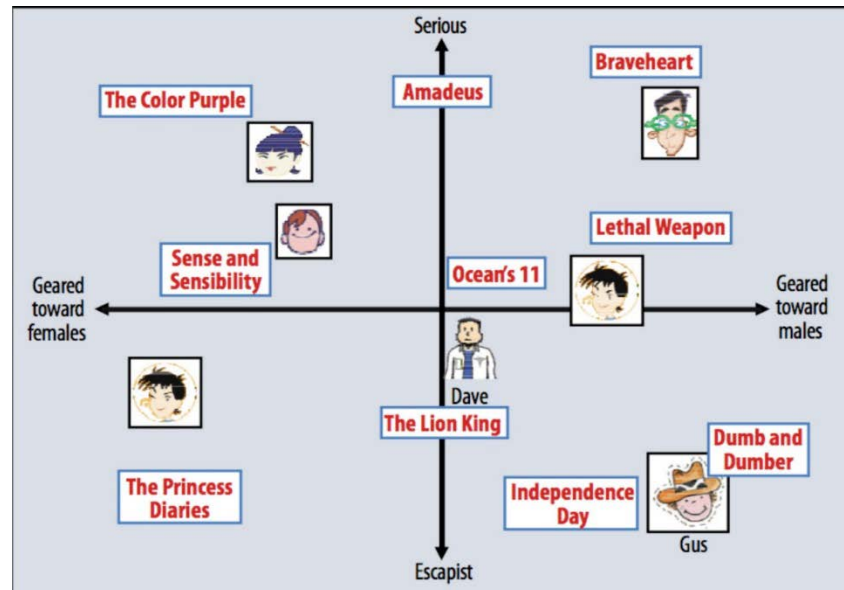
★★★★☆ (53)



★★★★☆ (33)

A Matrix Representation View

Users	Movie1	Movie2	Movie3	Movie4	Movie5	Movie6	...
User1	?	?	4	?	1	?	...
User2	2	5	2	?	?	2	...
User3	?	?	5	3	2	4	...
User4	1	?	?	4	?	?	...
User5	2	3	?	?	?	?	...
...



Summary: Methods to Learn

- Part I: Basics of data mining
 - Logistic regression, Naïve Bayes, K-Means, mixture models, ANN
- Part II: Text mining
 - Topic Models: pLSA, LDA
 - Word embedding
- Part III: Graphs/Network mining
 - Random walk, label propagation, spectral clustering, graph embedding, knowledge graph embedding
- Part IV: Recommender systems
 - Collaborative filtering, matrix factorization, Bayesian personalized ranking
 - Factorization machine, neural collaborative filtering
 - Recommendation as link prediction

Where to Find References? DBLP, CiteSeer, Google

- Data mining and KDD (SIGKDD: CDROM)
 - Conferences: ACM-SIGKDD, IEEE-ICDM, SIAM-DM, PKDD, PAKDD, etc.
 - Journal: Data Mining and Knowledge Discovery, KDD Explorations, ACM TKDD
- Database systems (SIGMOD: ACM SIGMOD Anthology—CD ROM)
 - Conferences: ACM-SIGMOD, ACM-PODS, VLDB, IEEE-ICDE, EDBT, ICDT, DASFAA
 - Journals: IEEE-TKDE, ACM-TODS/TOIS, JIIS, J. ACM, VLDB J., Info. Sys., etc.
- AI & Machine Learning
 - Conferences: AAI, IJCAI, COLT (Learning Theory), CVPR, ICML, NIPS, ICLR etc.
 - Journals: Machine Learning, Artificial Intelligence, Knowledge and Information Systems, IEEE-PAMI, etc.
- NLP
 - ACL, EMNLP, NAACL, etc.
- Web and IR
 - Conferences: SIGIR, WWW, CIKM, etc.
 - Journals: WWW: Internet and Web Information Systems,
- Statistics
 - Conferences: Joint Stat. Meeting, etc.
 - Journals: Annals of statistics, etc.
- Visualization
 - Conference proceedings: CHI, ACM-SIGGraph, etc.
 - Journals: IEEE Trans. visualization and computer graphics, etc.

Textbook

- Recommended: Jiawei Han, Micheline Kamber, and Jian Pei. [Data Mining: Concepts and Techniques](#), 3rd edition, Morgan Kaufmann, 2011
- References
 - "Data Mining: The Textbook" by Charu Aggarwal (<http://www.charuaggarwal.net/Data-Mining.htm>)
 - "Neural Networks and Deep Learning" (<https://www.springer.com/us/book/9783319944623>)
 - "Data Mining" by Pang-Ning Tan, Michael Steinbach, and Vipin Kumar (<http://www-users.cs.umn.edu/~kumar/dmbook/index.php>)
 - "Machine Learning" by Tom Mitchell (<http://www.cs.cmu.edu/~tom/mlbook.html>)
 - "Introduction to Machine Learning" by Ethem ALPAYDIN (<http://www.cmpe.boun.edu.tr/~ethem/i2ml/>)
 - "Pattern Classification" by Richard O. Duda, Peter E. Hart, David G. Stork (<http://www.wiley.com/WileyCDA/WileyTitle/productCd-0471056693.html>)
 - "The Elements of Statistical Learning: Data Mining, Inference, and Prediction" by Trevor Hastie, Robert Tibshirani, and Jerome Friedman (<http://www-stat.stanford.edu/~tibs/ElemStatLearn/>)
 - "Pattern Recognition and Machine Learning" by Christopher M. Bishop (<http://research.microsoft.com/en-us/um/people/cmbishop/prml/>)

slido



Audience Q&A Session

ⓘ Start presenting to display the audience questions on this slide.