CS-E4600
Algorithmic methods of data mining

Aristides Gionis
Dept of Computer Science

Slide set 2: Introduction to data mining

reading assignment

RLU book, chapter 1

what is data mining?

not a definite and clear answer

one potential definition:

use of efficient techniques to analyze very large collections of data and extract useful and possibly unexpected patterns

why need data mining?

huge amounts of data!

terabytes of data generated every second

mobile devices, digital photographs, web documents

facebook updates, tweets, blogs, user-generated content

transactions, sensor data, surveillance data

queries, clicks, browsing

cheap storage has made possible to store this data

need to analyze the raw data to extract knowledge

why need data mining?

data is power!

large amounts of data can be more powerful than complex algorithms and hand-crafted models

> Google has solved many Natural Language Processing problems, simply by looking at large document collections
>
> simple example: misspellings, synonyms

today, data is one of the biggest assets of a company

> query logs in Google
>
> friendships and posts in Facebook
>
> tweets and followers in Twitter
>
> purchase transactions in Amazon

need to design ways to harness the collective intelligence

data is complex

different types: tables, documents time series, images, graphs, etc.

spatial and temporal aspects

interconnected data of different types

example: from the mobile phone we can collect:

user location

friendship information

check-ins to venues

opinions through twitter

images though cameras

queries to search engines

example 1: transaction data

millions of customers

data associated to customers (e.g., loyalty cards)

walmart: 20 million transactions per day

AT&T: 300 million calls per day

credit card companies: billions of transactions per day

applications:

detect fraudulent usage of credit cards

bank evaluates whether to give a loan

personalized advertising and offers

example 2: document data

view the web as a document repository:

~50 billions of webpages

indexed by Google in Aug 2018

wikipedia:

~5.7 million articles in September 2018

online news portals:

thousands of new articles every day

twitter:

~500 million tweets every day (2017)

example 3: network data

the web graph:

50 billion pages linked via hyperlinks (2018)

facebook friendship graph:

~2.2 billion users (2018)

twitter follower graph:

~500 million users (2018)

instant messenger:

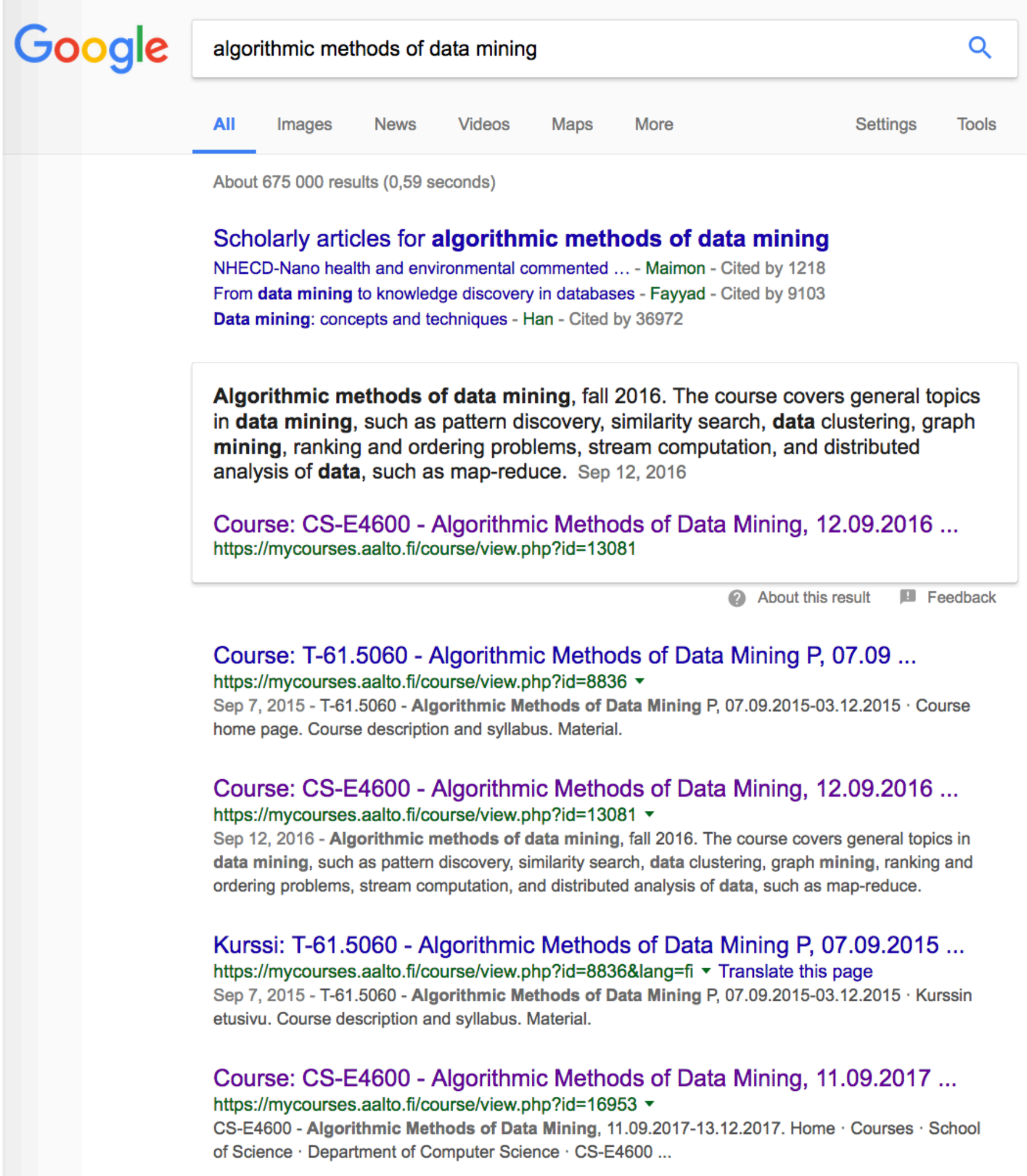
~1billion users

blogs:

250 million blogs worldwide

document and network data

application: web-search and document ranking



example 4: genomic sequences

http://www.1000genomes.org/page.php

full sequence of 1000 individuals

3 billion nucleotides per person

lots more data:

medical history of the persons

gene expression data

example 5: environmental data

spatiotemporal data

climate data (just an example) https://www.ncdc.noaa.gov/

> "National Climatic Data Center (NCDC) is responsible for preserving, monitoring, assessing, and providing public access to the Nation's treasure of climate and historical weather data and information."

example 6: behavioral data

mobile phones record lots of user behavior information

- GPS position, and location-based social behavior
- camera photos, posted in social media
- communication via phone, SMS, or online social networks

Amazon collects items browsed or purchased, wishlist, etc.

Google records queries, page visits, clicks, etc.

data collected for millions of users on a daily basis

data types

relational

transaction

0-1 data

real-valued data

sequences

time series

graphs / networks

relational data

attributes

objects

id	buying price	maint price	doors	# of persons	lug boot	safety	acceptability
1	high	med	3	4	med	low	acc
2	high	high	3	more	big	low	unacc
3	high	vhigh	3	4	small	low	unacc
4	vhigh	low	4	more	med	med	acc
5	vhigh	low	5more	2	big	high	unacc
6	med	low	2	4	big	high	vgood
7	low	low	5more	4	small	high	good

relational data

collection of data objects and their attributes

attributes

also known as variables, features

objects

also known as records, points, samples, entities, instances

size: number of objects

dimensionality: number of attributes

sparsity: number of populated entries

attribute types

categorical

values in a set

ordinal (order but no obvious distance)

e.g., {high, med, low}

nominal (no order or comparison)

e.g., {red, blue, yellow}

numerical

e.g., temperature, time, length, count

discrete vs. continuous

mixed

real-valued data

relational data with numerical attributes

id	popul	pct urban	med income	below pov line	pct divorced	pct police	violent crime per popul
1	131	25	24	4	30	2	4
2	45	34	21	9	45	3	7
3	64	73	19	5	31	2	3
4	254	65	18	17	52	6	12
5	379	78	31	5	41	2	5
6	32	57	29	12	39	1	1
7	11	81	26	8	44	4	6

real-valued data

data can be seen as points in a multidimensional space

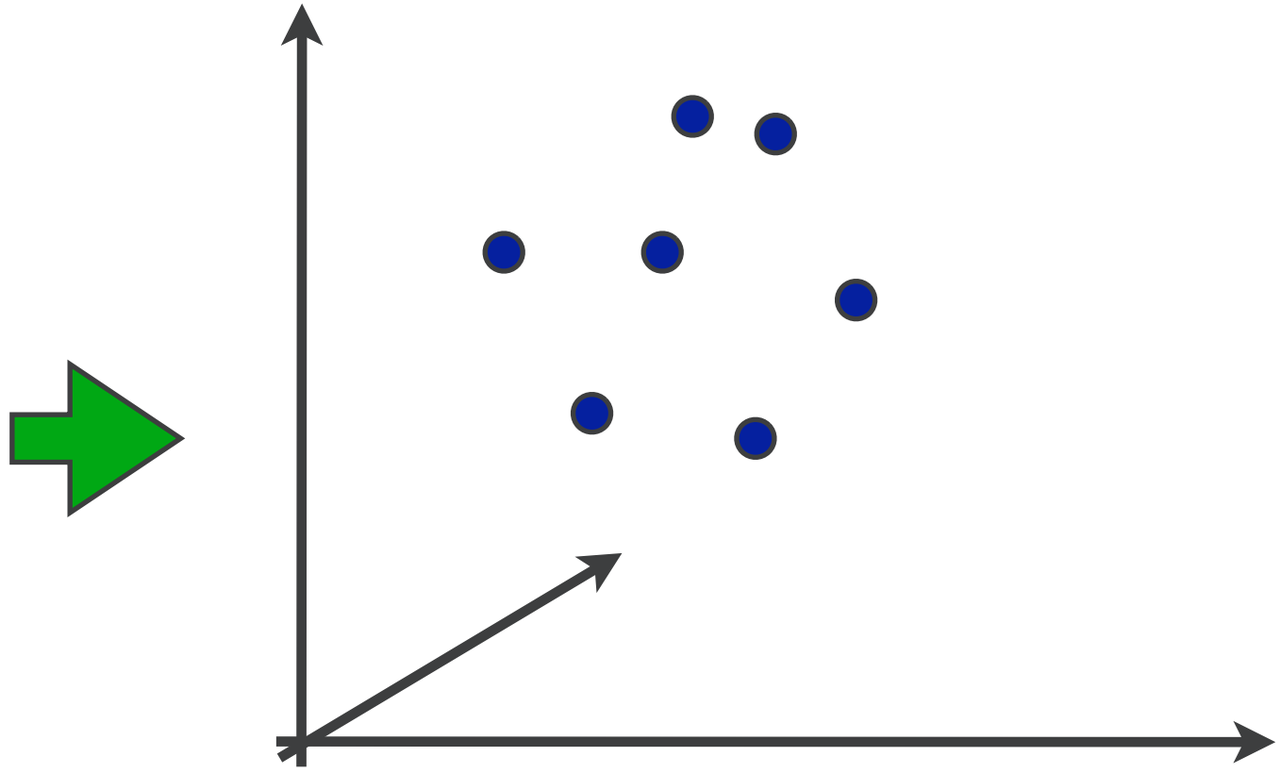
vector-space representation

each dimension represents a distinct attribute

number of points = number of objects

dimensionality = number of attributes

id	popul	pct urban	med income	below pov	pct divorced	pct police	violent crime per popul
1	131	25	24	4	30	2	4
2	45	34	21	9	45	3	7
3	64	73	19	5	31	2	3
4	254	65	18	17	52	6	12
5	379	78	31	5	41	2	5
6	32	57	29	12	39	1	1
7	11	81	26	8	44	4	6



0/1 data

relational data with binary attributes

student id	DB	OS	Algo	DS	DM	ML	AI
1	0	0	1	0	1	1	0
2	1	1	0	0	1	0	0
3	1	0	1	1	0	1	0
4	0	0	0	1	1	1	1
5	1	0	1	0	0	0	1
6	1	1	0	1	0	0	0
7	0	0	1	1	1	0	1

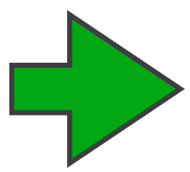
0/1 data

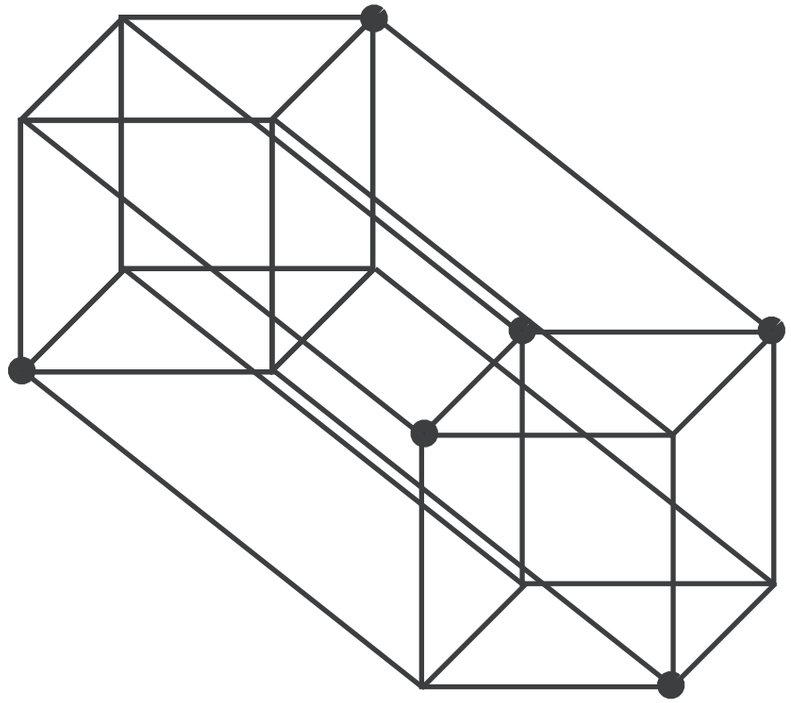
geometric interpretation still meaningful

data points map to the vertices of the hypercube

very difficult to visualize

student id	DB	OS	Algo	DS	DM	ML	AI
1	0	0	1	0	1	1	0
2	1	1	0	0	1	0	0
3	1	0	1	1	0	1	0
4	0	0	0	1	1	1	1
5	1	0	1	0	0	0	1
6	1	1	0	1	0	0	0
7	0	0	1	1	1	0	1





transaction data

student id	Courses
1	Algo, DM, ML
2	DB, OS, DM,
3	DB, Algo, DS, ML
4	Algo, DM, ML, AI
5	DB, Algo, AI
6	DB, OS, DS
7	Algo, DS, DM, AI

transaction data = 0/1 data

student id	Courses
1	Algo, DM, ML
2	DB, OS, DM,
3	DB, Algo, DS, ML
4	Algo, DM, ML, AI
5	DB, Algo, AI
6	DB, OS, DS
7	Algo, DS, DM, AI

student id	DB	OS	Algo	DS	DM	ML	AI
1	0	0	1	0	1	1	0
2	1	1	0	0	1	0	0
3	1	0	1	1	0	1	0
4	0	0	0	1	1	1	1
5	1	0	1	0	0	0	1
6	1	1	0	1	0	0	0
7	0	0	1	1	1	0	1

document data

consider a collection of documents

each document is represented by the set of words / terms it contains

bag-of-words representation — no ordering, grammar, syntax!

doc id	terms
1	model, phase, quantum, transition
2	dark, higgs, matter, quantum
3	higgs, model, phase, transition
4	dark, lattice, matter

document data

can be represented as transaction data or 0/1 data

doc id	terms
1	model, phase, quantum, transition
2	dark, higgs, matter, quantum
3	higgs, model, phase, transition
4	dark, lattice, matter

doc id	dark	higgs	lattice	matter	model	phase	quantum	transition
1	0	0	0	0	1	1	1	1
2	1	1	0	1	0	0	1	0
3	0	1	0	0	1	1	0	1
4	1	0	1	1	0	0	0	0

document data

can also consider how many times a term appears in each document

doc id	dark	higgs	lattice	matter	model	phase	quantum	transition
1	0	0	0	0	2	1	1	1
2	3	1	0	3	0	0	5	0
3	0	4	0	0	1	6	0	5
4	3	0	2	4	0	0	0	0

vector-space representation is more appropriate

transaction data is a general and useful abstraction

object A (of type X) is associated with object B (type Y)

transaction or 0/1 data representation

examples:

document contains term

student took class

user accessed web site

food contains ingredient

costumer bought product

person watched movie

X

Y

student id	DB	OS	Algo	DS	DM	ML	AI
1	0	0	1	0	1	1	0
2	1	1	0	0	1	0	0
3	1	0	1	1	0	1	0
4	0	0	0	1	1	1	1
5	1	0	1	0	0	0	1
6	1	1	0	1	0	0	0
7	0	0	1	1	1	0	1

when values are involved

object A (of type X) is associated with object B (type Y) with value C (numeric)

real-valued data or vector-space representation

examples:

document contains term k times

user reviewed movie with score x

compound contains element at fraction f

Y

X

id	popul	pct urban	med income	below pov line	pct divorced	pct police	violent crime per popul
1	131	25	24	4	30	2	4
2	45	34	21	9	45	3	7
3	64	73	19	5	31	2	3
4	254	65	18	17	52	6	12
5	379	78	31	5	41	2	5
6	32	57	29	12	39	1	1
7	11	81	26	8	44	4	6

transaction data is a general abstraction

object A (of type X) is associated with object B (type Y)

X and Y can be the same type

examples:

page links to page

user is a friend of user

person called person

protein interacts with protein

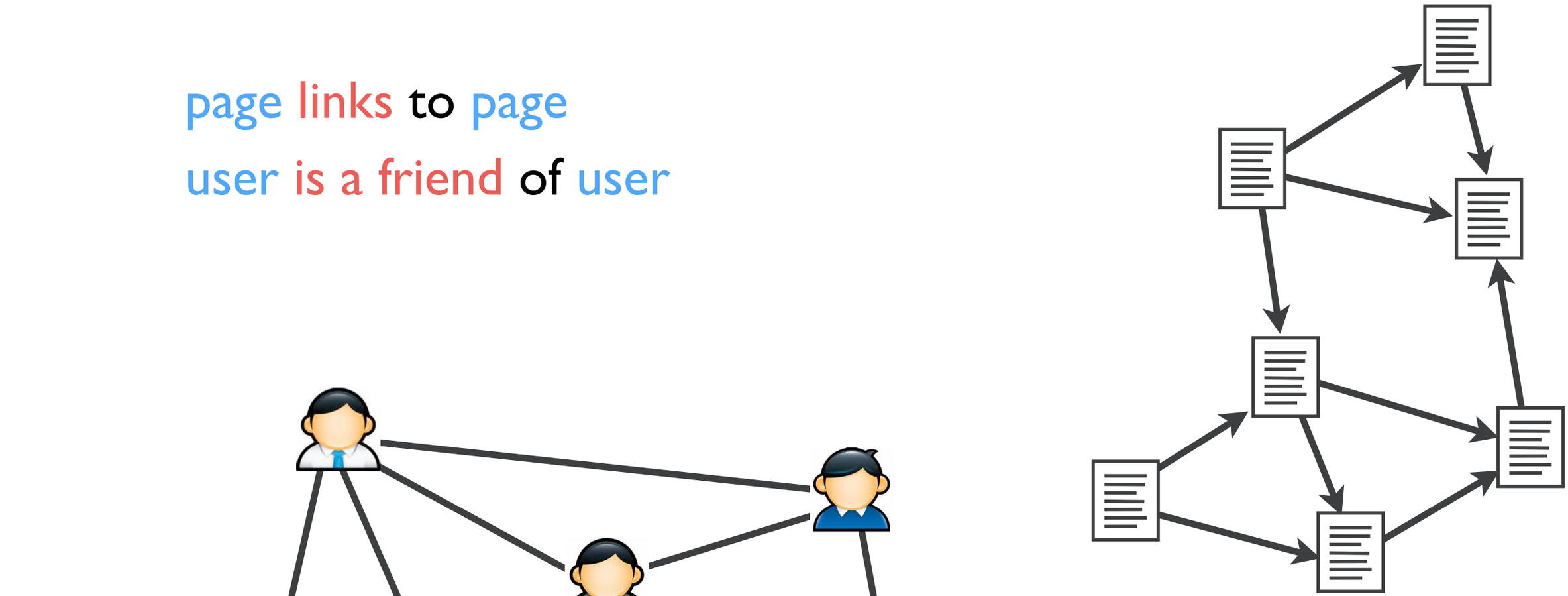
X

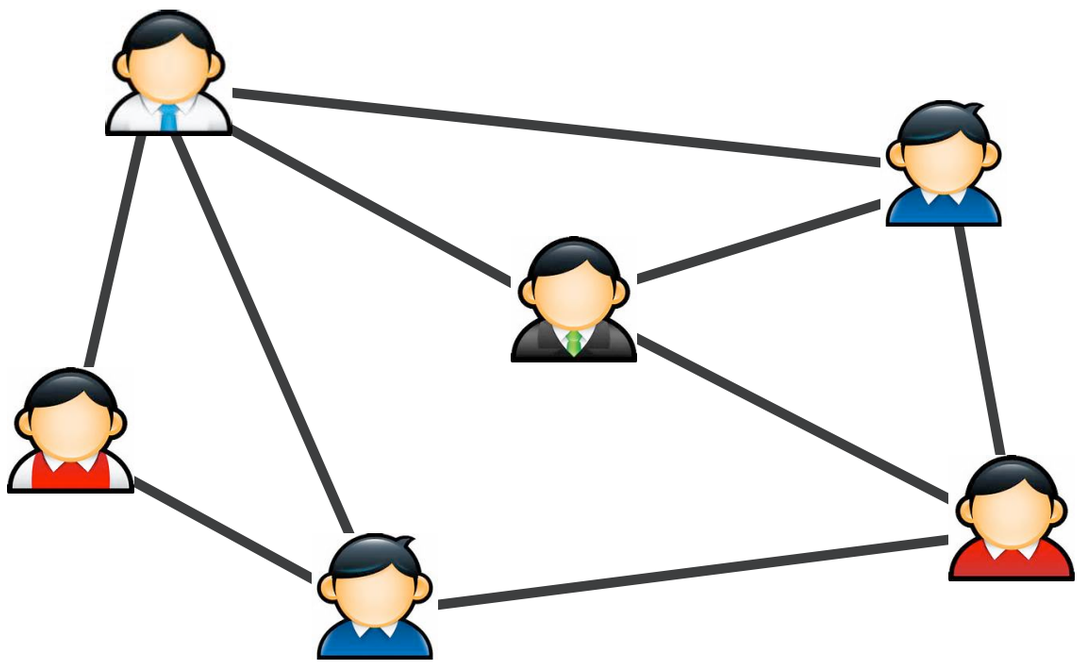
	1	2	3	4	5	6	7
1	0	0	1	0	1	1	0
2	1	1	0	0	1	0	0
3	1	0	1	1	0	1	0
4	0	0	0	1	1	1	1
5	1	0	1	0	0	0	1
6	1	1	0	1	0	0	0
7	0	0	1	1	1	0	1

graph data

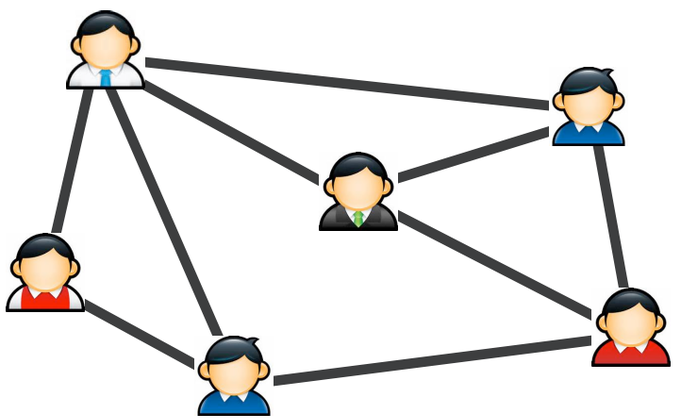
page links to page

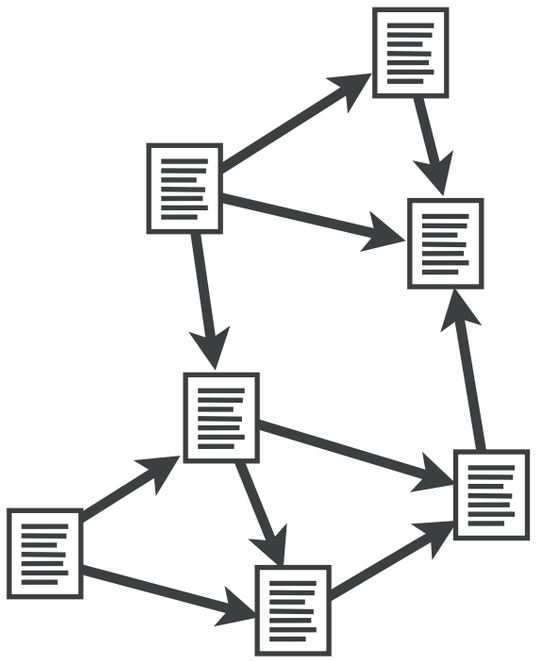
user is a friend of user





graph data = transaction data = 0/1 data





X

	1	2	3	4	5	6	7
1	0	0	1	0	1	1	0
2	1	1	0	0	1	0	0
3	1	0	1	1	0	1	0
4	0	0	0	1	1	1	1
5	1	0	1	0	0	0	1
6	1	1	0	1	0	0	0
7	0	0	1	1	1	0	1

X

undirected vs. directed graph?

symmetric data matrix

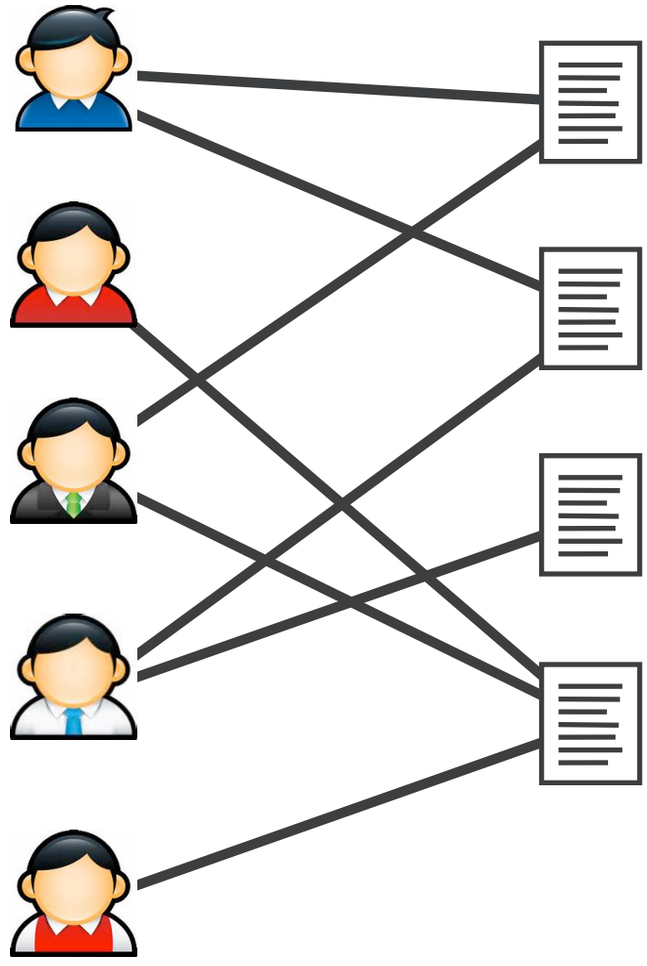
graph interpretation of transaction data with X ≠ Y ?

Y

X

student id	DB	OS	Algo	DS	DM	ML	AI
1	0	0	1	0	1	1	0
2	1	1	0	0	1	0	0
3	1	0	1	1	0	1	0
4	0	0	0	1	1	1	1
5	1	0	1	0	0	0	1
6	1	1	0	1	0	0	0
7	0	0	1	1	1	0	1

a bipartite graph!



X Y

ordered data

genomic sequence data

```
GGTTCCGCCTTCAGCCCCGCGCC
CGCAGGGCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG
```

data is a long ordered string

or many strings

ordered data

time series

sequence of temporally ordered numerical values

summary : data types

relational data

objects and attributes

quite general, categorical, numerical, or mixed attributes

real-valued data or vector-space data points

transaction data = 0/1 data = set data

graph data (can also be seen as 0/1 data)

ordered data (strings or time series)

what would you do if…

imagine you have access to the amazon.com data

who has bought what

what information you would extract from the data and how you would use it?

what would you do if…

imagine you have access to the Google data

who has queried about what and looked at which page

what information you would extract from the data and how you would use it?

what would you do if…

imagine you have access to the Bloomberg data

how stocks fluctuate over time

what information you would extract from the data and how you would use it?

in-class discussion

discuss with your neighbor

what kind of analysis you would do with that data

pick one dataset

google, amazon, bloomberg, or other

analytics, knowledge discovery, or data-driven application