Data Mining and Analysis

Data Privacy

CSCE 676 :: Fall 2019
Texas A&M University
Department of Computer Science & Engineering
Prof. James Caverlee

Resources

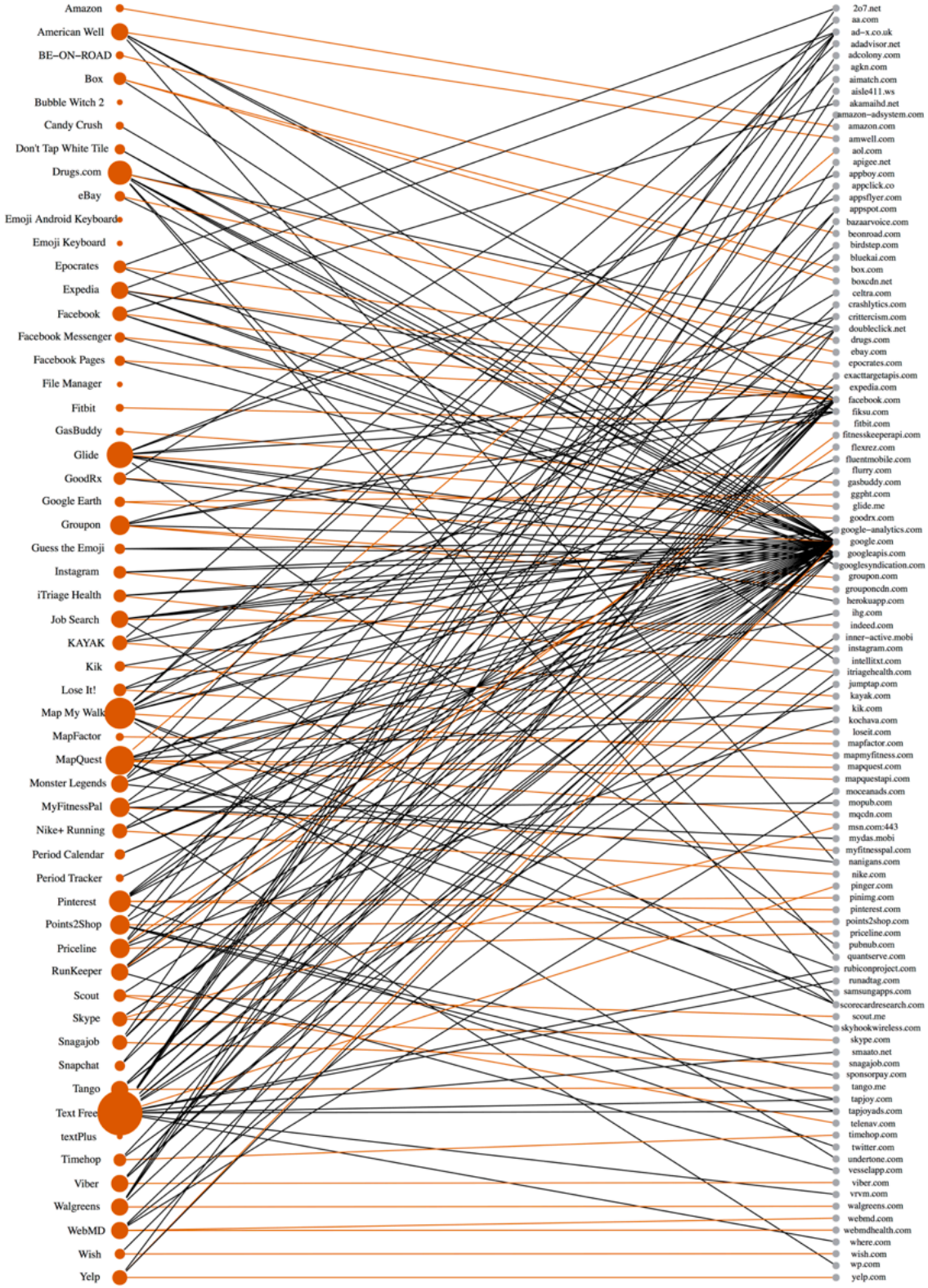
Li Xiong Data Privacy and Security course @Emory

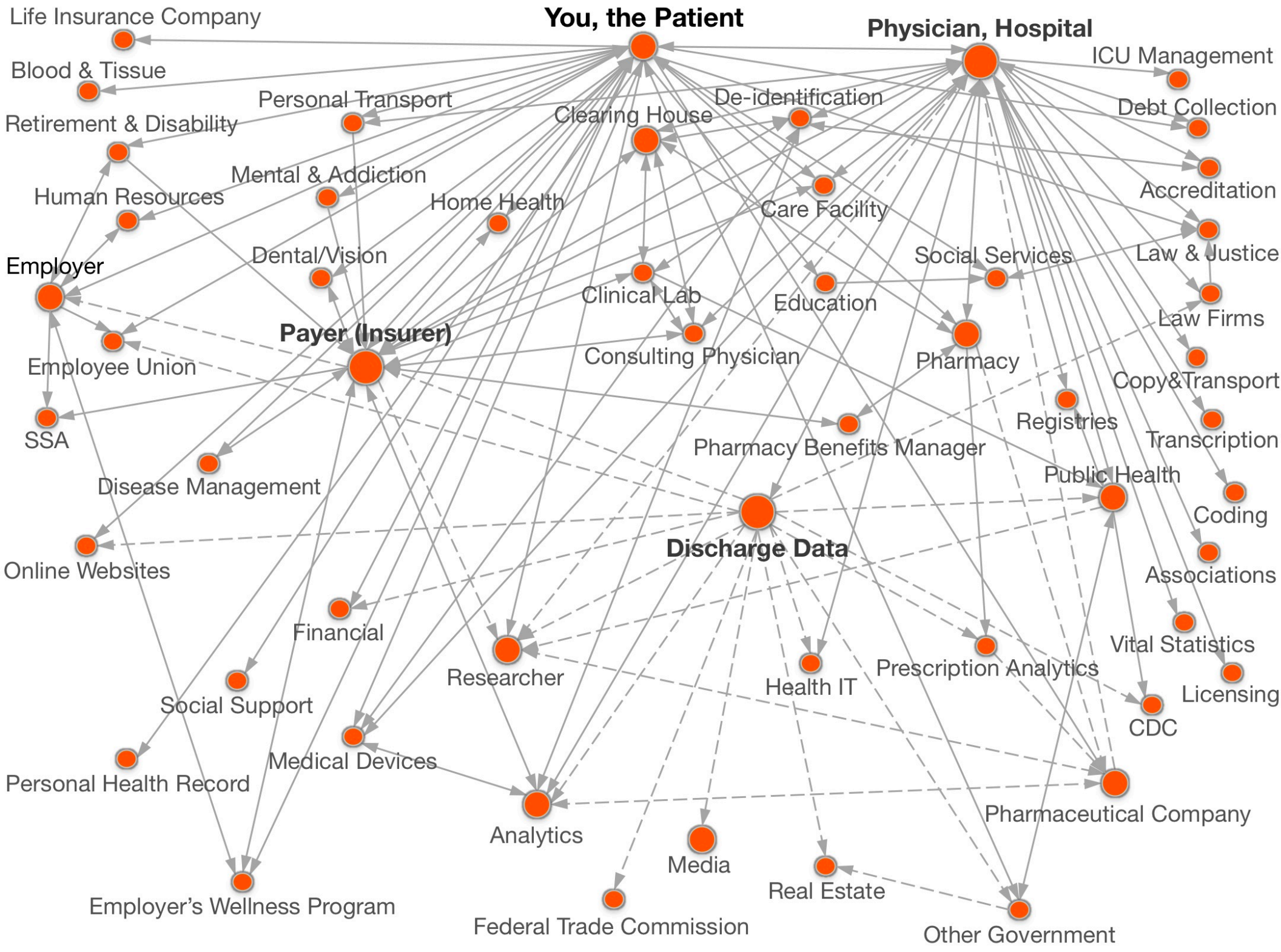
http://www.mathcs.emory.edu/~lxiong/cs573_f16/schedule.html

Privacy-preserving data publishing: A survey of recent developments; k-anonymity (Sweeney)

A firm foundation for private data analysis, CACM, 2011; Designing statistical privacy for your data, CACM, 2015

Differential Privacy in the Wild. SIGMOD 2017 Tutorial. Machanavajjhala, Hay, and He.


Amazon
American Well
BE–ON–ROAD
Box
Bubble Witch 2
Candy Crush
Don't Tap White Tile
Drugs.com
eBay
Emoji Android Keyboard
Emoji Keyboard
Epocrates
Expedia
Facebook
Facebook Messenger
Facebook Pages
File Manager
Fitbit
GasBuddy
Glide
GoodRx
Google Earth
Groupon
Guess the Emoji
Instagram
iTriage Health
Job Search
KAYAK
Kik
Lose It!
Map My Walk
MapFactor
MapQuest
Monster Legends
MyFitnessPal
Nike+ Running
Period Calendar
Period Tracker
Pinterest
Points2Shop
Priceline
RunKeeper
Scout
Skype
Snagajob
Snapchat
Tango
Text Free
textPlus
Timehop
Viber
Walgreens
WebMD
Wish
Yelp
2o7.net
aa.com
ad–x.co.uk
adadvisor.net
adcolony.com
agkn.com
aimatch.com
aisle411.ws
akamaihd.net
amazon–adsystem.com
amazon.com
amwell.com
aol.com
apigee.net
appboy.com
appclick.co
appsflyer.com
appspot.com
bazaarvoice.com
beonroad.com
birdstep.com
bluekai.com
box.com
boxcdn.net
celtra.com
crashlytics.com
crittercism.com
doubleclick.net
drugs.com
ebay.com
epocrates.com
exacttargetapis.com
expedia.com
facebook.com
fiksu.com
fitbit.com
fitnesskeeperapi.com
flexrez.com
fluentmobile.com
flurry.com
gasbuddy.com
ggpht.com
glide.me
goodrx.com
google–analytics.com
google.com
googleapis.com
googlesyndication.com
groupon.com
grouponcdn.com
herokuapp.com
ihg.com
indeed.com
inner–active.mobi
instagram.com
intellitxt.com
itriagehealth.com
jumptap.com
kayak.com
kik.com
kochava.com
loseit.com
mapfactor.com
mapmyfitness.com
mapquest.com
mapquestapi.com
moceanads.com
mopub.com
mqcdn.com
msn.com:443
mydas.mobi
myfitnesspal.com
nanigans.com
nike.com
pinger.com
pinimg.com
pinterest.com
points2shop.com
priceline.com
pubnub.com
quantserve.com
rubiconproject.com
runadtag.com
samsungapps.com
scorecardresearch.com
scout.me
skyhookwireless.com
skype.com
smaato.net
snagajob.com
sponsorpay.com
tango.me
tapjoy.com
tapjoyads.com
telenav.com
timehop.com
twitter.com
undertone.com
vesselapp.com
viber.com
vrvm.com
walgreens.com
webmd.com
webmdhealth.com
where.com
wish.com
wp.com
yelp.com


Life Insurance Company
You, the Patient
Physician, Hospital
ICU Management
Blood & Tissue
Personal Transport
De-identification
Debt Collection
Retirement & Disability
Clearing House
Mental & Addiction
Accreditation
Human Resources
Home Health
Care Facility
Law & Justice
Employer
Dental/Vision
Social Services
Clinical Lab
Education
Law Firms
Payer (Insurer)
Consulting Physician
Pharmacy
Employee Union
Copy&Transport
Registries
SSA
Transcription
Pharmacy Benefits Manager
Public Health
Disease Management
Coding
Discharge Data
Online Websites
Associations
Financial
Vital Statistics
Prescription Analytics
Researcher
Health IT
Licensing
Social Support
CDC
Medical Devices
Personal Health Record
Pharmaceutical Company
Analytics
Media
Real Estate
Employer's Wellness Program
Federal Trade Commission
Other Government

Waze is an awesome driving app that also lets hackers stalk you

Researchers at the University of California-Santa Barbara recently discovered a Waze vulnerability that allowed them to create thousands of "ghost drivers" that can monitor the drivers around them—an exploit that could be used to track Waze users in real-time. They proved it to me by tracking my own movements around San Francisco and Las Vegas over a three-day period.

PLEASE ROB ME
Raising awareness about over-sharing
Check out our guest blog post on the CDT website.
Like
Share
31,360 people like this.
Check your own Twitter timeline for checkins
Are you curious if people can see your checkins?
Enter your Twitter username and find out.
Your Twitter username
Find!
Share
More Info
Home
Why
Made Possible By
Foursquare
Twitter
@boyvanamstel
@frankgroeneveld
@barryborsboom

Health Information Privacy

HIPAA

- Understanding HIPAA Privacy
- HIPAA Administrative Simplification Statute and Rules
- Enforcement Activities & Results
- How to File a Complaint
- News Archive
- Frequently Asked Questions

PSQIA

- Understanding PSQIA Confidentiality
- PSQIA Statute & Rule
- Enforcement Activities & Results
- How to File a Complaint

Health Information Privacy

The Office for Civil Rights enforces the HIPAA Privacy Rule, which protects the privacy of individually identifiable health information; the HIPAA Security Rule, which sets national standards for the security of electronic protected health information; the HIPAA Breach Notification Rule, which requires covered entities and business associates to provide notification following a breach of unsecured protected health information; and the confidentiality provisions of the Patient Safety Rule, which protect identifiable information being used to analyze patient safety events and improve patient safety.

The Health Insurance Portability and Accountability Act of 1996 (HIPAA) Privacy, Security and Breach Notification Rules

Learn about the Rules' protection of individually identifiable health information, the rights granted to individuals, breach notification requirements, OCR's enforcement activities, and how to file a complaint with OCR.

Health Social Networks

CureTogether

Already a member? Sign in

The smarter way to find the best treatments.

Get access to millions of ratings comparing the real-world performance of treatments across 637 health conditions.

treatment | member ratings (num votes)

- sleep (489)
- Ibuprofen (Advil, Motrin) (472)
- Caffeine (451)
- Imitrex/migran (Sumatriptan) (430)
- Excedrin Migraine (361)
- Naproxen (Aleve) (360)
- Acetaminophen/Paracetamol (Tylenol) (355)
- Dietary Changes (350)
- Eliminate caffeine (coffee, tea, sodas) (324)
- Topiramate (Topamax) (322)
- Neck Massage (320)

Sign up - it's anonymous and free.

Enter your email (no spam, we're here to do good)

will not be publicly displayed

Choose a password

6 characters or more

Sign up

CureTogether is now part of 23andMe

learn more about your genetics

MedHelp it's your health™

Login | Free Membership | Connect

Forums | Health Tools | Health Information | Experts | My MedHelp

Search MedHelp

The World's Largest Health Community

Helping millions of people be their healthiest

MedHelp empowers more than 12 million people each month to take control of their health and find answers to their medical questions.

LEARN MORE

TODAY'S PULSE

Timeline: Your Body, Cigarette-Free

When you quit smoking, your body begins to restore itself almost immediately. Find out how your health improves in the moments, hours, months and years after you put out your last cigarette.

New Perspectives in Parkinson's Disease

Parkinson's disease affects over 1 million Americans. This Missouri Medicine report sheds light on new developments in treating the disease.

HEALTH STAT OF THE WEEK

How are you sleeping this week?

On average, we sleep 41 minutes more on the weekends than during the week.

SUN MON TUE WED THU FRI SAT

patientslikeme®

Username or Email | ●●●●●●●● | Login

Remember me | Forgot password?

"Because of PatientsLikeMe, we are better able to recognize warning signs... [and] keep things in perspective. In short, PatientsLikeMe empowers us."

"PatientsLikeMe has provided me with new friends- people who are experiencing the same problems as I am."

"I joined because I didn't want to feel alone anymore. Simply put. And I knew that I could be helpful [by sharing] my experience."

Your experience matters.

At PatientsLikeMe, you'll find the resources and support you need to improve your own health. Ask questions, get answers, and help make healthcare better for everyone.

Your experience matters.

164,460 patients

1,000+ conditions

Join Now! (It's free)

Search this site | Search

Find Patients like you

Explore our Treatment reports

Learn about Symptoms

Check for your Conditions

DailyStrength

Home | Support Groups | Health Blogs | Expert Answers | Treatments | People | Find a Cond

MEMBER SIGN IN

DailyStrength | Connect

Email or Username

Password

Remember me

Login

Lost your password?

No account?

Register now. It's FREE, quick, and easy.

Join Now

Become a member now and join the largest, most comprehensive network of people sharing their knowledge, experiences and support.

What our members love about DailyStrength:

- Making friends, journaling, tracking goals, sharing photos and Hugs
- Active discussion boards, expert advice, treatment reviews
- Privacy tools to keep personal content protected

Popular Support Groups | View All »

ADHD	Deep Vein Thrombosis	Loneliness
Alcoholism	Depression	Menopause
Anxiety	Diabetes	Migraines
Asthma	Dieting	Miscarriage
Autism	Eating Disorders	Multiple Sclerosis

Virgin sacks 13 over Facebook 'chav' remarks

Ben Quinn

@BenQuinn75

Fri 31 Oct 2008 20.01 EDT

Virgin Atlantic has sacked 13 cabin crew after they posted messages on Facebook referring to passengers as "chavs" and making jokes about faulty engines.

The airline said yesterday the employees' behaviour was "totally inappropriate" and "brought the company into disrepute".

Disciplinary action was launched last week amid claims flight crew had criticised its safety standards and insulted passengers on the social networking site.

Netflix Prize 2: (Privacy) Apocalypse Now?

Netflix passed out a $1 million check for its Netflix Prize yesterday and …

by **Nate Anderson** - Sept 22 2009, 3:30pm CDT

Netflix yesterday passed out a $1 million check to the stats geeks of "BellKor's Pragmatic Chaos," the team which won the Netflix Prize by improving on the company's algorithms for picking the movies that Netflix subscribers might like to watch next. Chief Product Officer Neil Hunt then announced a sequel, the Netflix Prize 2, but one law professor is already calling the new contest a "multi-million dollar privacy blunder."

The Netflix data set released for use with the first prize was supposedly "anonymized," but security researchers found a way to link the anonymous movie recommendations with data from other sites in order to identify individual Netflix users. If your information appeared in the first data set, someone knowing a bit about your movie preferences could figure out the complete set of movies that you rated on Netflix—an information leak that would certainly not please all users.

For the Netflix Prize 2, the company is releasing even more information, despite the security weaknesses of the first data set. Gender, ZIP Codes, ages, genre ratings, and more will be released in a data set with more than 100 million entries. That has law professor Paul Ohm seeing a bright Netflix red.

Ohm has worried for years about privacy breaches, and his newest paper argues that "anonymizing" data doesn't actually keep data anonymous. Given the tremendous advances in "reidentification" technology over the last few years, Ohm argues that the release of so much new data should actually constitute a "privacy breach."

"Researchers have known for more than a decade that gender plus ZIP code plus birthdate uniquely identifies a significant percentage of Americans," he wrote in a blog entry. "True, Netflix plans to release age not birthdate, but simple arithmetic shows that for many people in the country, gender plus ZIP code plus age will narrow their private movie preferences down to at most a few hundred people. Netflix needs to understand the concept of 'information entropy': even if it is not revealing information tied to a single person, it is revealing information tied to so few that we should consider this a privacy breach."

Because Netflix already knows about the privacy implications of its data releases, it can't simply claim ignorance, and in Ohm's view the company might well be liable for damages.

Netflix Prize 2 crowdsourcing contest killed due to privacy concerns

Netflix nixes contest following the filing of a lawsuit and FTC discussions

By Alpha Doggs on Fri, 03/12/10 - 4:46pm.

2 Comments Print Share g+1 Like 1 More

Netflix announced on Friday that it is killing a contest announced last September designed to offer the public big money for helping the company come up with a better way to gauge movie watchers' preferences.

From Chief Product Officer Neil Hunt, on the company blog:

"In the past few months, the Federal Trade Commission (FTC) asked us how a Netflix Prize sequel might affect Netflix members' privacy, and a lawsuit was filed by KamberLaw LLC pertaining to the sequel. With both the FTC and the plaintiffs' lawyers, we've had very productive discussions centered on our commitment to protecting our members' privacy.

We have reached an understanding with the FTC and have settled the lawsuit with plaintiffs. The resolution to both matters involves certain parameters for how we use Netflix data in any future research programs.

In light of all this, we have decided to not pursue the Netflix Prize sequel that we announced on August 6, 2009."

AT&T Labs-Research, Yahoo Research and other members of the Bellkor's Pragmatic Chaos team in September celebrated their win in the inaugural 3-year-long Netflix Prize contest. They earned $1 million in beating out 40,000 other teams from 186 countries to improve upon Netflix's Cinematch system for predicting which movies customers will like or dislike.

No sooner did Netflix hand out its prizes than it announced Netflix Prize 2, which was to reward winners at 6 and 18 months.

But the new contest immediately raised privacy concerns among some. Blogger Paul Ohm (an associate professor of Law at the University of Colorado Law School) said he respected the effort Netflix went to to protect customer data handed to contestants in Netflix Prize 1, but had serious privacy concerns about Netflix Prize 2, which involved dishing out demographic, gender and other data.

NETFLIX

Why 'Anonymous' Data Sometimes Isn't

By Bruce Schneier 12.13.07

Last year, Netflix published 10 million movie rankings by 500,000 customers, as part of a challenge for people to come up with better recommendation systems than the one the company was using.

The New York Times

Business Day

Technolo

WORLD | U.S. | N.Y. / REGION | BUSINESS | TECHNOLOGY | SCIENCE | HE

Marketers Can Glean Private Data on Facebook

TARGET

TECH | 2/16/2012 @ 11:02AM | 837,678 views

How Target Figured Out A Teen Girl Was Pregnant Before Her Father Did

Privacy vs. Confidentiality

Confidentiality

Prevent disclosure of information to unauthorized users

Privacy

Prevent disclosure of personal information to unauthorized users

Control of how personal information is collected and used

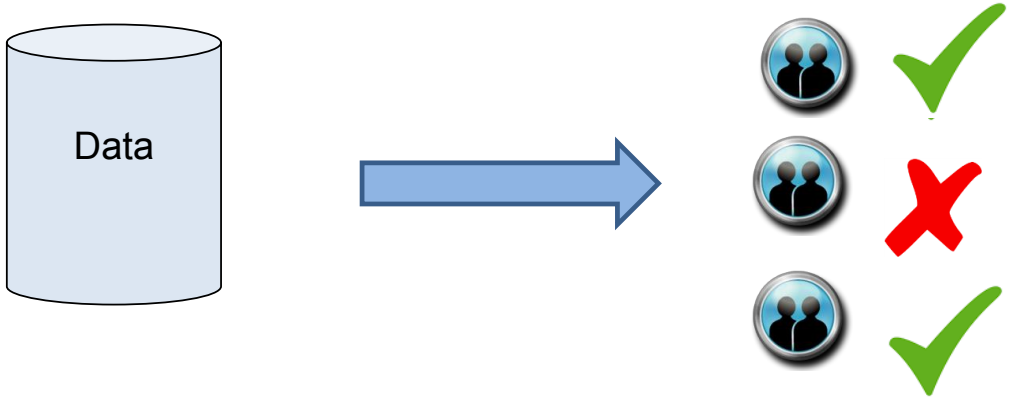
Prevent identification of individuals

Data Privacy and Security Measures

- Access control
  - Restrict access to the (subset or view of) data to authorized users
- Cryptography
  - Use encryption to encode information so it can be only read by authorized users (protected in transmit and storage)
- Inference control
  - Restrict inference from accessible data to sensitive (non-accessible) data

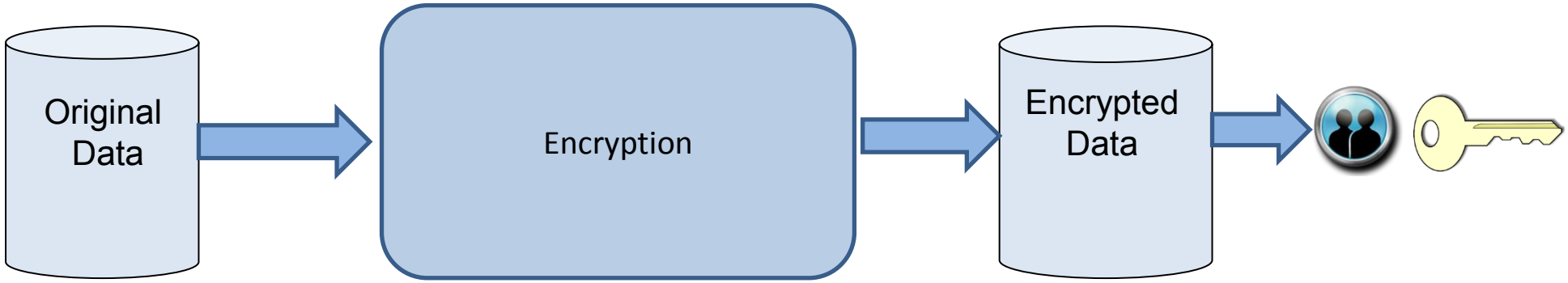
Access Control

Selective restriction of access to the data to authorized users



Cryptography

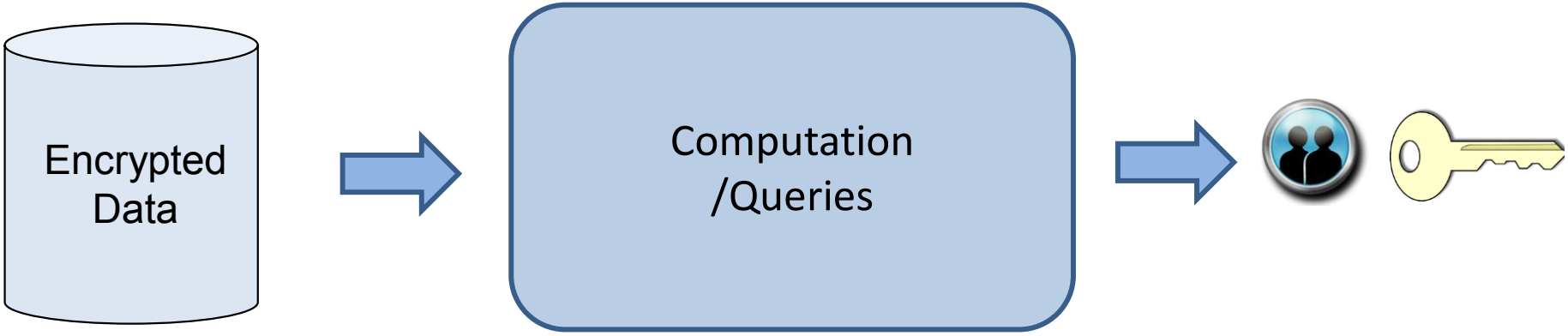
Encoding data in a way that only authorized users can read it



Applications of Crypto

Secure data outsourcing

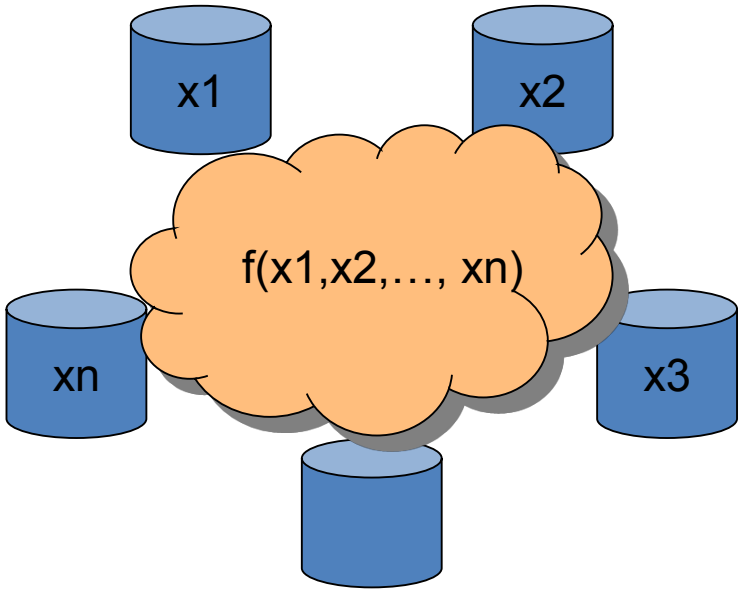
Support computation and queries on encrypted data



Applications of Crypto

Multi-party secure computation (aka secure function evaluation)

Securely compute a function without revealing private inputs

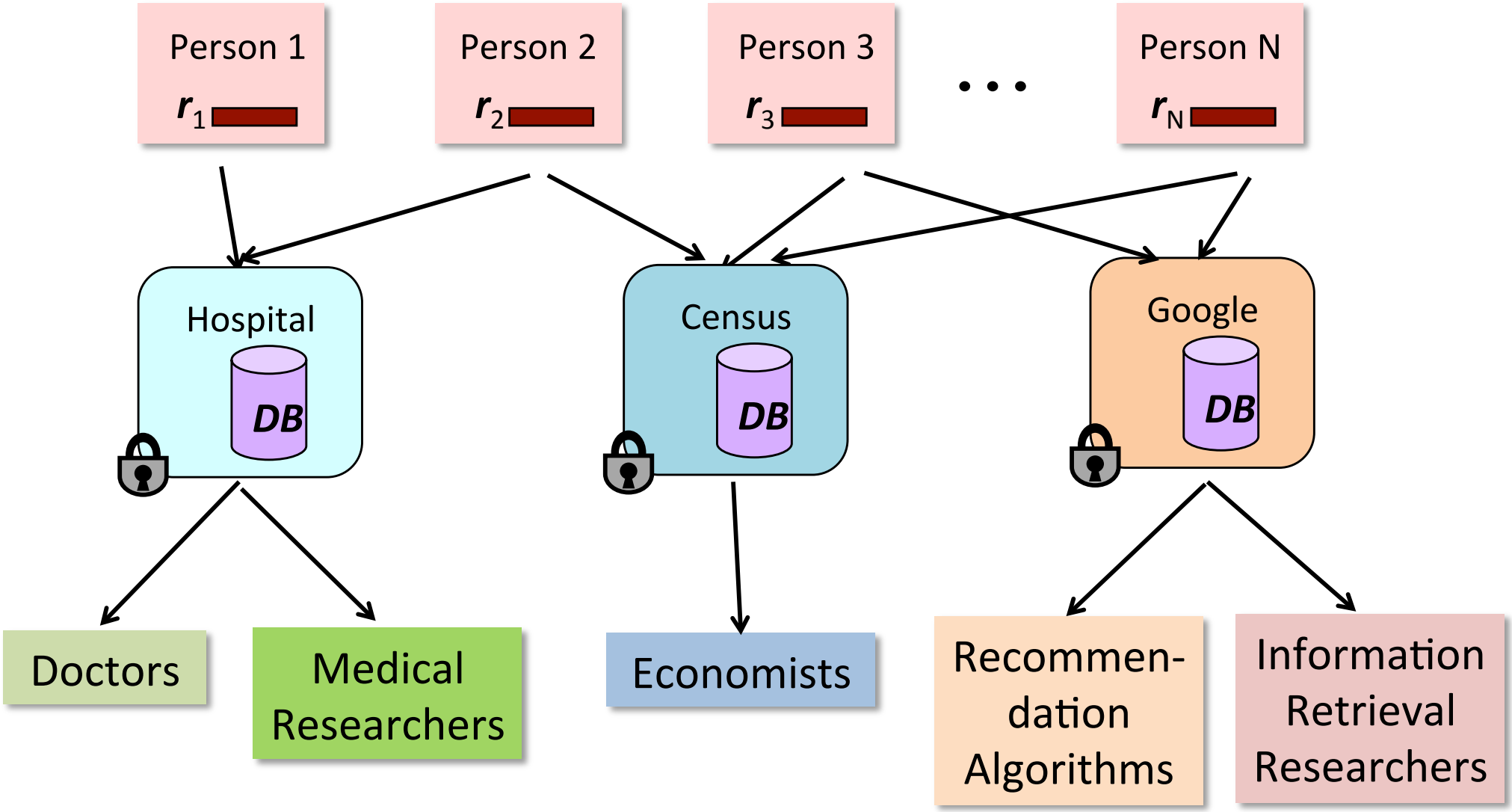


Inference Control

Prevent inference from accessible information to individual information (not accessible)

Technologies

- Privacy-preserving data mining (typically tied tightly to specific algorithm)
- De-identification and Anonymization (input perturbation)
- Differential Privacy (output perturbation)


Person 1
r_1
Person 2
r_2
Person 3
r_3
...
Person N
r_N
Hospital
DB
Census
DB
Google
DB
Doctors
Medical Researchers
Economists
Recommen-dation Algorithms
Information Retrieval Researchers

Privacy-Preserving Data Mining

Privacy-preserving Data Mining

Privacy-preserving association rule mining

Privacy-preserving k-means clustering

Privacy-preserving decision trees

Privacy-preserving …

…

…

https://scholar.google.com/scholar?start=0&q=privacy+preserving&hl=en

privacy preserving

Scholar About 2,180,000 results (**0.04** sec)

Privacy-preserving data mining
R Agrawal, R Srikant - ACM Sigmod Record, 2000 - dl.acm.org
A fruitful direction for future data mining research will be the development of techniques that incorporate **privacy** concerns. Specifically, we address the following question. Since the primary task in data mining is the development of models about aggregated data, can we …
Cited by 3944 Related articles All 31 versions

Privacy preserving data mining
Y Lindell, B Pinkas - Annual International Cryptology Conference, 2000 - Springer
In this paper we introduce the concept of **privacy preserving** data mining. In our model, two parties owning confidential databases wish to run a data mining algorithm on the union of their databases, without revealing any unnecessary information. This problem has many …
Cited by 1378 Related articles All 34 versions

Revealing information while **preserving privacy**
I Dinur, K Nissim - Proceedings of the twenty-second ACM SIGMOD …, 2003 - dl.acm.org
We examine the tradeoff between **privacy** and usability of statistical databases. We model a statistical database by an n-bit string d 1,.., dn, with a query being a subset q⊆[n] to be answered by Σ iεq d i. Our main result is a polynomial reconstruction algorithm of data from …
Cited by 808 Related articles All 15 versions

Privacy-preserving set operations

Secure Sum

Multiple parties, each holds a value: we want to find the sum V

(know it lies somewhere in [0…n]

Lead site 1 generates random number R in [0…n]

Sends $R+v_1$ mod n to site 2

Secure Sum

Site l receives:

$$V = R + \sum_{j=1}^{l-1} v_j \bmod n.$$

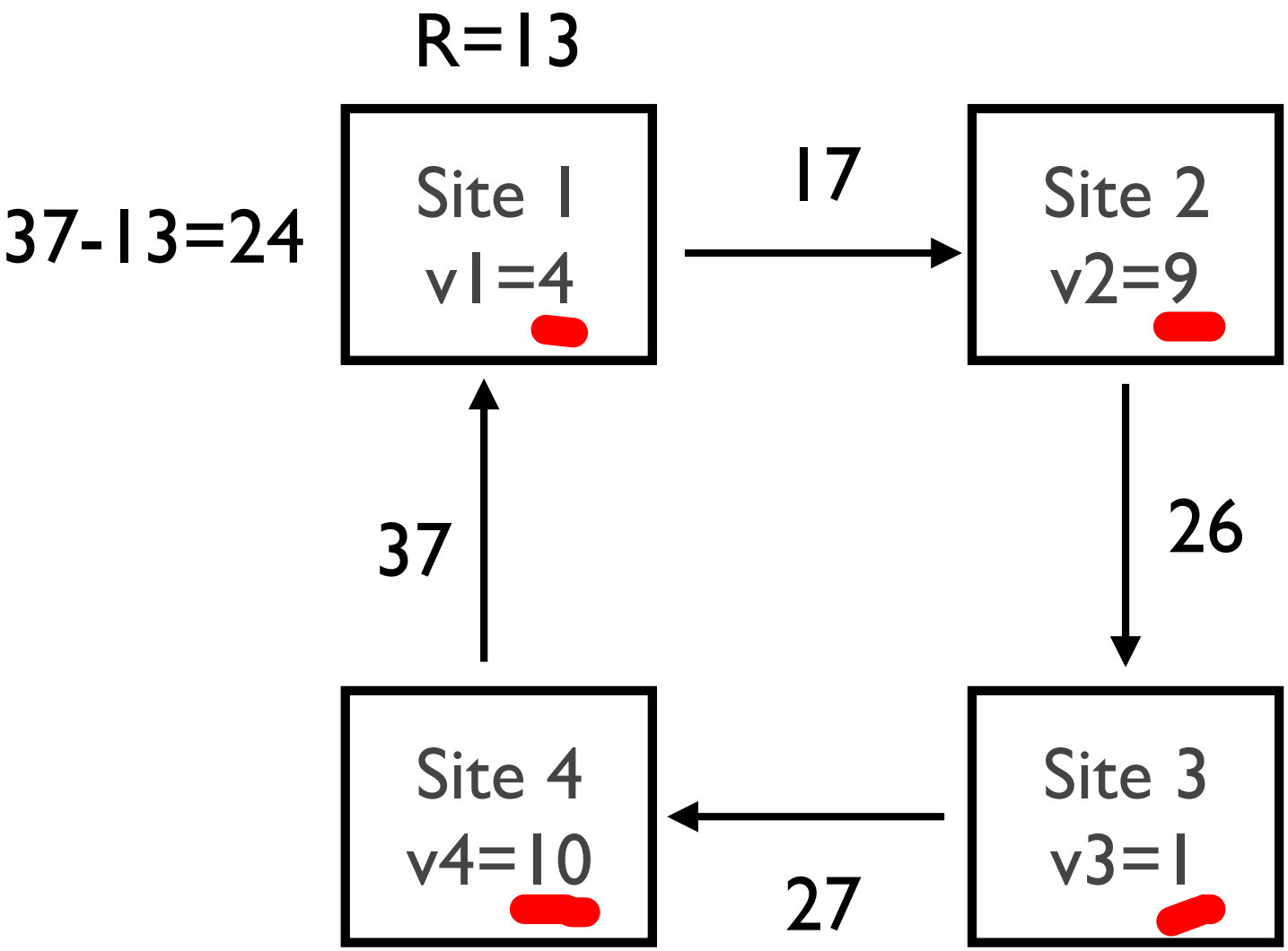
Then computes and passes along:

$$R + \sum_{j=1}^{l} v_j \bmod n = (v_j + V) \bmod n$$

Finally, site 1 can subtract R to find V

Secure Sum: Example

Answer = 24



Secure Sum Issues?

Privacy-Preserving K-Means Clustering over Vertically Partitioned Data

Jaideep Vaidya
Department of Computer Sciences
Purdue University
250 N University St
West Lafayette, IN 47907-2066
jsvaidya@cs.purdue.edu

Chris Clifton
Department of Computer Sciences
Purdue University
250 N University St
West Lafayette, IN 47907-2066
clifton@cs.purdue.edu

ABSTRACT

Privacy and security concerns can prevent sharing of data, derailing data mining projects. Distributed knowledge discovery, if done correctly, can alleviate this problem. The key is to obtain valid results, while providing guarantees on the (non)disclosure of data. We present a method for k-means clustering when different sites contain different attributes for a common set of entities. Each site learns the cluster of each entity, but learns nothing about the attributes at other sites.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*Data mining*; H.2.7 [**Database Management**]: Database Administration—*Security, integrity, and protection*; H.2.4 [**Database Management**]: Systems—*Distributed databases*

Imagine the following scenario. A law enforcement agency wants to cluster individuals based on their financial transactions, and study the differences between the clusters and known money laundering operations. Knowing the differences and similarities between normal individuals and known money launderers would enable better direction of investigations. Currently, an individual's financial transactions may be divided between banks, credit card companies, tax collection agencies, etc. Each of these (presumably) has effective controls governing release of the information. These controls are not perfect, but violating them (either technologically or through insider misuse) reveals only a subset of an individual's financial records. The law enforcement agency could promise to provide effective controls, but overcoming these controls now gives access to an individual's entire financial history. This raises justifiable concerns among privacy advocates.

Privacy and data mining can coexist. The problem with

Recall: k-means

Input is **k** (the number of clusters), **data points** in Euclidean space

0. Initialize clusters by picking one point per cluster

Loop:

1. Place each point in the cluster whose current centroid is nearest
2. Find the new centroid for each cluster

Privacy-Preserving k-means

Goal: Cluster the known set of common entities without revealing any value that the clustering is based on.

Input:

Each “site” provides one attribute of all items.

Output:

Assignment of entities to clusters.

Cluster centers themselves.

Data is “vertically partitioned”

	Price	Color	Weight	…
Item 1	200	Blue	900g	…
Item 2	700	Red	300g	…
Item 3	200	Black	800g	…
Item 4	500	Yellow	500g	…
Item 5	800	Red	900g	…
Item 6	300	Yellow	300g	…
Item 7	500	Black	900g	…
⋮	⋮	⋮	⋮	

“site 1” “site 2”

r: # of users, each having different attributes for the same set of items.

r parties, k clusters, n points.

```
repeat
    for all j = 1...r do
        for i = 1...k do
            μ_ij ← μ'_ij
            Cluster[i] = ∅
        end for
    end for
    for g = 1...n do
        for all j = 1...r do
            {Compute the distance vector X⃗_j (to each cluster) for point g.}
            for i = 1...k do
                x_ij = |data_gj −_D μ_ij|
            end for
        end for
        Each site puts g into Cluster[closest_cluster]
        {closest_cluster is Algorithm 3}
    end for
    for all j = 1...r do
        for i = 1...k do
            μ'_ij ← mean of j's attributes for points in Cluster[i]
        end for
    end for
until checkThreshold {Algorithm 2}
```

Securely finding the closest cluster

The problem is formally defined as follows. Consider r parties $P_1, \ldots, P_r$, each with their own k-element vector $\vec{X}_i$:

$$P_1 \text{ has } \vec{X}_1 = \begin{bmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{k1} \end{bmatrix}, P_2 \text{ has } \begin{bmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{k2} \end{bmatrix}, \ldots, P_r \text{ has } \begin{bmatrix} x_{1r} \\ x_{2r} \\ \vdots \\ x_{kr} \end{bmatrix}.$$

The goal is to compute the index l that represents the row with the minimum sum. Formally, find

$$\underset{i=1..k}{argmin}\left(\sum_{j=1..r} x_{ij}\right)$$

Securely finding the closest cluster

The security of the algorithm is based on three key ideas.

- Disguise the site components of the distance with random values that cancel out when combined.
- Permute the order of clusters so the real meaning of the comparison results is unknown.
- Compare distances so only the comparison result is learned; no party knows the distances being compared.

{Stage 1: Between P_1 and all other parties}

P_1 generates r random vectors $\vec{V_i}$ summing to $\vec{0}$ (see Algorithm 4).

P_1 generates a random permutation π over k elements

for all $i = 2 \ldots r$ **do**

  $\vec{T_i}$ (at P_i) $= add_and_permute(\vec{V_i}, \pi(\text{at } P_1), \vec{X_i}(\text{at } P_i))$ {This is the permutation algorithm described in Section 2.2}

end for

P_1 computes $\vec{T_1} = \pi(\vec{X_1} + \vec{V_1})$

{Stage 2: Between all but P_2 and P_r}

for all $i = 1, 3 \ldots r - 1$ **do**

    P_i sends $\vec{T_i}$ to P_r

end for

P_r computes $\vec{Y} = \vec{T_1} + \sum_{i=3}^{r} \vec{T_i}$

{Stage 3: Involves only P_2 and P_r}
$minimal \leftarrow 1$
for j=2..k **do**
  if $secure_add_and_compare(Y_j + T_{2j} < Y_{minimal} + T_{2minimal})$ **then**
    $minimal \leftarrow j$
  end if
end for

{Stage 4: Between P_r (or P_2) and P_1}
Party P_r sends $minimal$ to P_1
P_1 broadcasts the result $\pi^{-1}(minimal)$

Check Threshold

Algorithm 2 checkThreshold: Find out if the new means are sufficiently close to old means

Require: Th is a threshold for termination, Random number generator *rand* produces values uniformly distributed over $0..n-1$ spanning (at least) twice the domain of the distance function $-_D$.

for all $j = 1 \ldots r$ **do**
  $d_j \leftarrow 0$
  for $i = 1 \ldots k$ **do**
    $d_j \leftarrow d_j + |\mu'_{ij} -_D \mu_{ij}|$
  end for
end for
{Securely compute if $\sum d_j \leq Th$.}
At P_1: $m = \text{rand}()$
for j=1 ... r-1 **do**
  P_i sends $m + d_j \pmod n$ to P_{j+1}
end for
At P_r: $m = m + d_r$
At P_1: $Th' = Th + r$
P_1 and P_r return $secure_add_and_compare(m - Th' \pmod n > Th' - m \pmod n)$ {Secure comparison is described in Section 2.3.}

Issues for PP Data Mining

Methods are often tied closely to specific data mining algorithms, so hard to generalize

(though some common sharing of primitives like “secure sum”)

Privacy efforts can be nullified by strategic collusive behavior

…