Data Security and Privacy

Yennun Huang

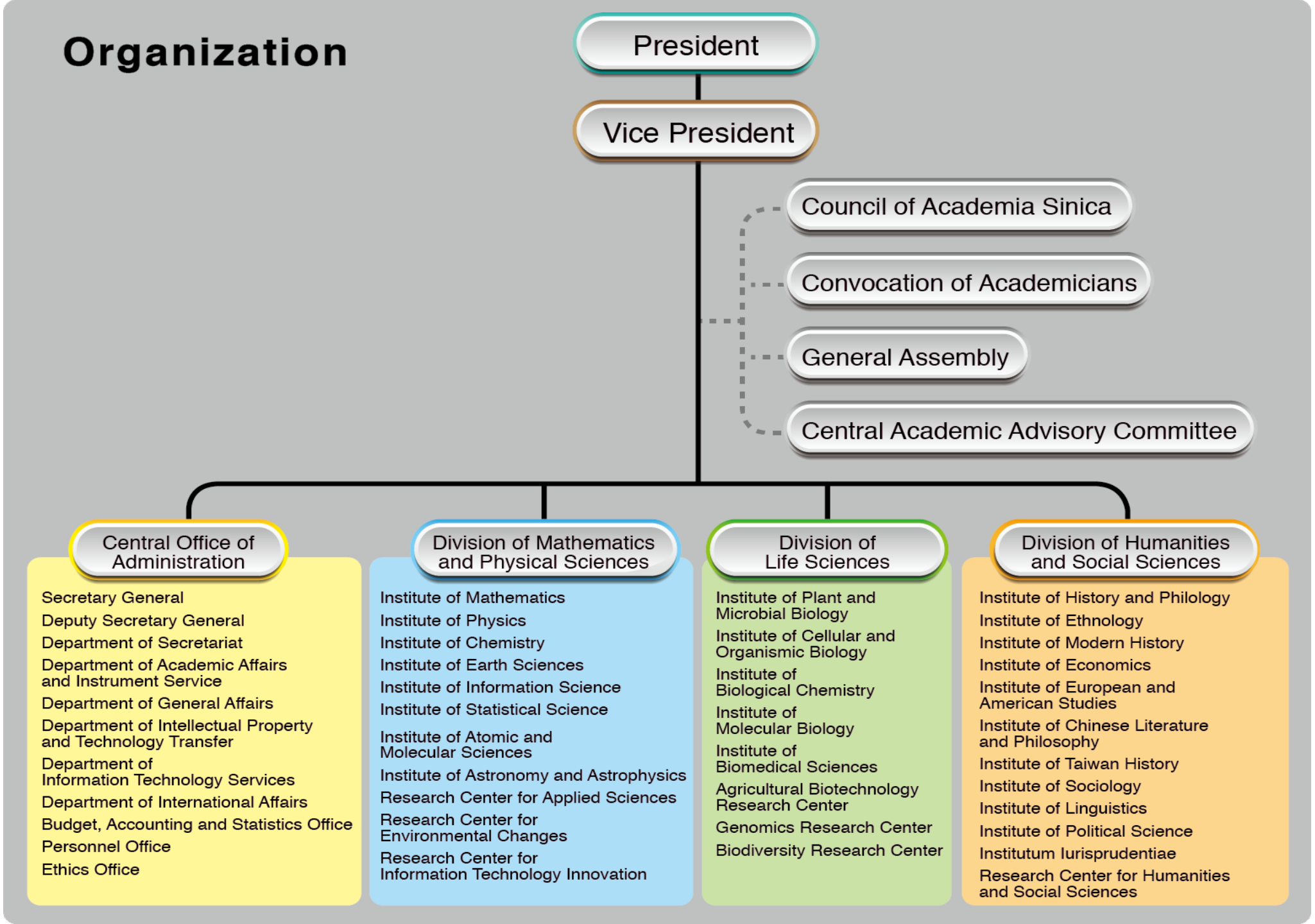
Research Center for Information Technology Innovation

Academia Sinica, Taiwan

Academia Sinica

- The most preeminent academic institution in Taiwan
- Founded in Republic of China in 1928 by the Nationalist government.
- Directly under President Office
- Roughly 10% of the Taiwan National science and technology R&D budget goes to Academia Sinica each year (10 Billion NT dollars).
- Promote international cooperation and scholarly exchanges that will accelerate the overall development of academic research in Academia Sinica and Taiwan.

Organization



CTI - Three Thematic Centers

- Grid & Scientific Computing Thematic Center
  - Distributed Cloud Computing
- Taiwan Information Security Thematic Center
  - Headquarter of National TWISC centers
- Intelligent & Ubiquitous Computing Thematic Center
  - FinTech/RegTech
  - IoT platforms and applications

Grid & Scientific Computing Center

One of ten Tier-1 centers of the World-Wide Large Hadron Collider (LHC) Computing Grid (WLCG)

- Collaborate with the European Laboratory of Particle Physics (CERN) on development of key components of the Distributed Cloud Operating System (DiCOS): distributed job management, distributed data management and the integration.
- First DiCOS was released in Sep. 2013 supporting ATLAS and AMS

Data Security and Privacy

World's Biggest Data B | PRDC 2017 | Big Data, for better or

Secure https://www.sciencedaily.com/releases/2013/05/130522085217.htm

Apps Google World's Biggest Dat Imported From IE 歡迎來到 Facebook 蘋果日報 | Apple Dai 聯合新聞網：每個人都 自由時報電子報

SD Health Tech Enviro Society Quirky Store A- A A+ Search

Science News

from research organizations

Print Email Share

Big Data, for better or worse: 90% of world's data generated over last two years

Date: May 22, 2013

Source: SINTEF

Summary: A full 90 percent of all the data in the world has been generated over the last two years. Internet-based companies are awash with data that can be grouped and utilized. Is this a good thing?

Share:

RELATED TOPICS

Computers & Math

- Computers and Internet
- Information Technology
- Hacking

Science & Society

- Surveillance

FULL STORY

Related Stories

Next Generation of Statistical Tools to Be Developed for the Big Data Age

Sep. 21, 2016 — Statisticians are developing new ways to interpret the unprecedented amounts of data being generated continuously all around us. Whether it is smart phones packed full of sensors measuring your ... **read more »**

Innovations Are Needed If Big Data Is to Boost

World's Biggest Data Br

informationisbeautiful.net/visualizations/worlds-biggest-data-breaches-hacks

Google Facebook 聯合新聞網：觸 自由時報電子報 蘋果日報 | Apple 中時電子報 CTnews -

World's Biggest Data Breaches

Selected losses greater than 30,000 records

(updated 5th Jan 2017)

interesting story

YEAR *BUBBLE COLOUR* YEAR METHOD OF LEAK *BUBBLE SIZE* NO OF RECORDS STOLEN DATA SENSITIVITY HIDE FILTER

latest

2016

Brazzers ixSense Interpark Lynda.com Netflix Twitter account National Childbirth Trust yAsUGym Quest Diagnostics Red Cross Blood Service

Clinton campaign Dailymotion 85200000

Mail. ru 25, 000, 000

Philippines' Commissi on Electi

Mutuelle Generale de la Police Syrian Government

Cellebrite

Anthem 80, 000, 000

Friend Finder Network 412, 000, 000

MySpace 164, 000, 000

Privatization Agency of the Republic of Serbia

Banner Health

Linux Ubuntu

Filter by...

ORGANISATION

- all
- academic
- app
- banking
- energy
- financial
- gaming
- government
- healthcare
- legal
- media
- military
- retail
- tech
- telecoms
- transport
- web

METHOD OF LEAK

- all
- accidentally published
- configuration error
- hack
- hacked
- inside job
- leak
- lost / stolen computer
- lost / stolen media
- poor security
- vulnerability

Hacker Tries To Sell 427 Milllion Stolen MySpace Passwords For $2,800

May 27, 2016 // 01:20 PM EDT

Written by
LORENZO FRANCESCHI-BICCHIERAI
STAFF WRITER

There's an oft-repeated adage in the world of cybersecurity: There are two types of companies, those that have been hacked, and those that don't yet know they have been hacked.

MySpace, the social media behemoth that was, is apparently in the second category. The same hacker who was selling the data of more than 164 million LinkedIn users last week now claims to have 360 million emails and passwords of MySpace users, which would be one of the largest leaks of passwords ever. And it looks like the data

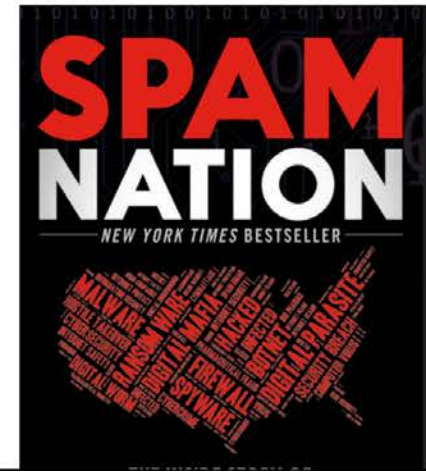
19 Online Cheating Site AshleyMadison Hacked

JUL 15

Large caches of data stolen from online cheating site **AshleyMadison.com** have been posted online by an individual or group that claims to have completely compromised the company's user databases, financial records and other proprietary information. The still-unfolding leak could be quite damaging to some 37 million users of the hookup service, whose slogan is "Life is short. Have an affair."

The data released by the hacker or hackers — which self-identify as **The Impact Team** — includes sensitive internal data stolen from Avid Life Media (ALM), the Toronto-based firm that owns AshleyMadison as well as related hookup sites **Cougar Life** and **Established Men**.

My New Book!



Data Breaches Events

- http://www.informationisbeautiful.net/visualizations/worlds-biggest-data-breaches-hacks/

Privacy Breach Events (1)

- In 2006, AOL released the Internet search terms that more than 650,000 of its subscribers entered over a three-month period,
  - substituted numeric IDs for the subscribers' real user names
- IN 2010, NetFlix included 100 million movie ratings, along with the date of the rating, a unique ID number for the subscriber, and the movie info.
  - Could identify targets by matching their Netflix reviews with data from other sites like IMDb.
  - Found that if you knew a few movies a Netflix subscriber had rented in a given time period, you could reverse-engineer the data and find out the rest of their viewing history.

Privacy Breach Events (2)

- The state of Massachusetts distributed a research dataset containing de-identified insurance reimbursement records of Massachusetts state employees that had been hospitalized. To protect the employees' privacy, their names were stripped from the dataset, but the employees' date of birth, zip code, and sex was preserved to allow for statistical analysis.
- Sweeney was able to re-identify the governor's records by searching for the "de-identified" records that matched the Governor's date of birth, zip code, and sex. She learned this information from the Cambridge voter registration list, which she purchased for $20. Sweeney then generalized her findings, arguing that up to 87% of the U.S. population could be uniquely identified by their 5-digit ZIP code, date of birth, and sex based on the 1990 census.

Data De-Identification is difficult

- Encryption

  How to analyze and compute on encrypted data?

- Anonymization

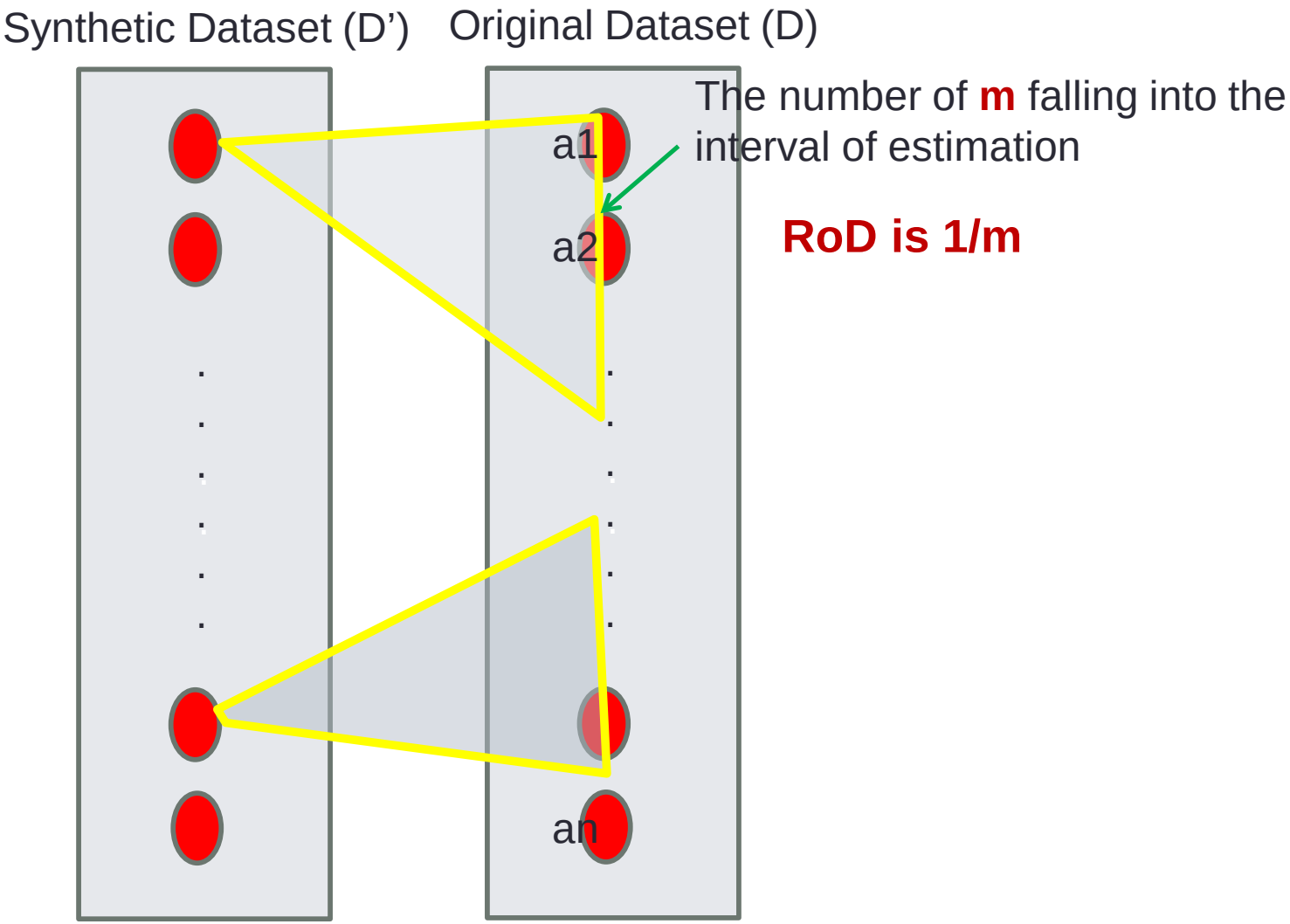
  Re-identification is possible

- Access mediation/control

  With multiple queries, re-identification is possible

- Adding noise
  - Permutation
  - Differential Privacy

Concept (RoD definition)



Evaluation on the RoD (numerical attributes)

- An example of evaluation on the RoD for the assigned sensitive attribute HIV using numerical attributes height and weight with the maximal magnitude of noise is 10 for height, 20 for weight and 1 for HIV.

- ✓ **Original identification Prob. =1**
- ✓ **RoD for conditions of 159cm and 60kg = 1/5**
- ✓ **RoD to the individual for the sensitive attribute HIV = 2/5**

Original Dataset (D)

Attributes		
Height (cm)	Weight (kg)	HIV (Y/N)
148	76	0
149	60	0
150	95	0
150	69	0
156	85	0
159	60	1
166	80	1
166	62	0
170	80	0
170	90	0

Synthetic Dataset (D')

Attributes		
Height (cm)	Weight (kg)	HIV (Y/N)
150	59	0
145	70	0
155	100	0
157	49	1
150	68	1
169	75	0
175	95	1
158	47	0
177	86	0
171	85	0

Estimate maximal noise (N_{Max}) and epsilon

- **Theorem 2.** *The privacy parameter* ε *is equivalent to* $-\frac{ln2}{MAX(Lap(1/\lambda))} \cdot ln(1-\gamma)$ *related to noise estimation* $MAX(Lap(1/\lambda))$ *and confidence probability* γ.

Proof: Let an actual value be ω and its corresponding synthetic value be $\varpi = \omega + Lap(1/\lambda)$. Let $\frac{\omega-\varpi}{\omega}$ be equivalent to the error rate $\pm\delta$ and $Lap(1/\lambda)$ be the Laplace noise with confidence probability γ, where γ can be defined as [12]:

$$\gamma = P[\omega - \delta\omega \leq \varpi \leq \omega + \delta\omega]. \tag{6}$$

Because $\varpi = \omega + Lap(1/\lambda)$, the Eq. (6) can be rewritten as:

$$\begin{aligned} \gamma &= P[-\delta\omega \leq Lap(1/\lambda) \leq \delta\omega] \\ &= P[Lap(1/\lambda) \leq |\delta\omega|]. \end{aligned} \tag{7}$$

Research on De-Identification

- Practical tools on Differential Privacy
- Evaluation of the privacy level on differential privacy
- Integration of various de-identification techniques
  - K-Anonymity, differential privacy, encryption
- Tools for linkage attacks