

CHƯƠNG 5

AN TOÀN CSDL THỐNG KÊ

Giảng viên: TS. Trần Thị Lượng

Mục tiêu

- Chúng ta đi sâu vào các vấn đề suy diễn trên các CSDL thống kê.
- Thảo luận một số kỹ thuật bảo vệ cơ bản:
 - **Kỹ thuật dựa vào khái niệm**
 - **Kỹ thuật dựa vào hạn chế**
 - **Kỹ thuật dựa vào gây nhiễu**
- Đánh giá chung về đặc trưng của các kỹ thuật này.

Tài liệu tham khảo

- *White Paper:*

- *Interference Attacks to Statistical Databases: Data Suppression, Concealing Controls and Other Security Trends* (Salvador Mandujano - Department of Computer Sciences- Purdue University -West Lafayette, IN USA 47907)
- *New Efficient Attacks on Statistical Disclosure Control Mechanisms* (Cynthia Dwork and Sergey Yekhanin- Microsoft Research)
- *OPTIMAL DISCLOSURE LIMITATION STRATEGY IN STATISTICAL DATABASES* (George T. Duncan¹ and Sumitra Mukherjee²)
- ...

Nội dung



Giới thiệu



Các khái niệm cơ bản



Một số kiểu tấn công suy diễn



Các kỹ thuật chống tấn công suy diễn

Tài liệu tham khảo

- **Website:**

- **Tổng cục thống kê – Việt Nam:**

- <http://www.gso.gov.vn/default.aspx?tabid=228&ItemID=1915>

- **SDB Liên hợp quốc (UNO)**

- <http://unstats.un.org/unsd/databases.htm>

- **SDB kinh tế khối Châu Âu (UNECE)**

- <http://w3.unece.org/pxweb/Dialog/>

- **SDB WTO**

- <http://stat.wto.org/Home/WSDBHome.aspx?Language=>

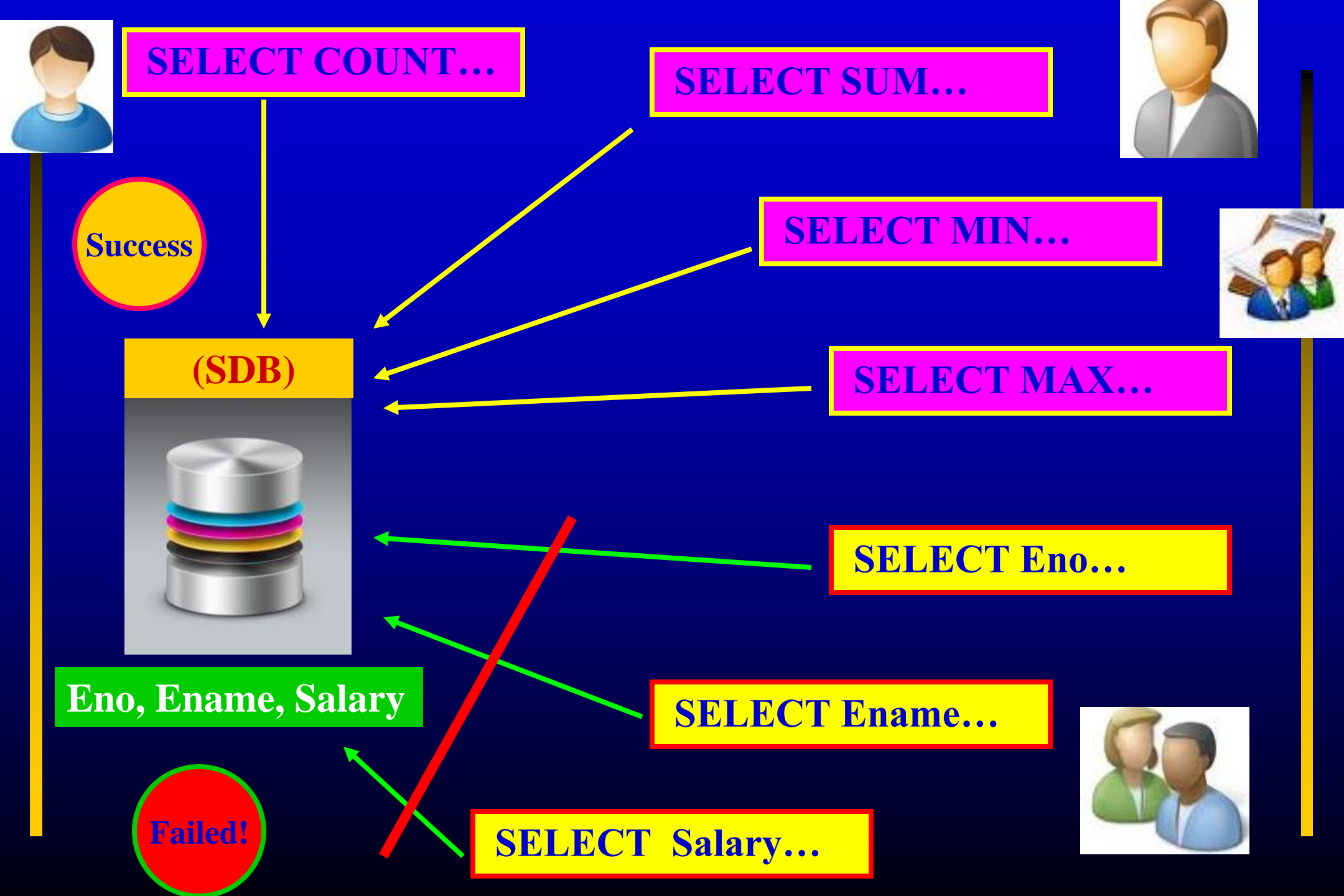
Question

- *CSDL thống kê là gì?*
- *CSDL thống kê khác CSDL quan hệ?*

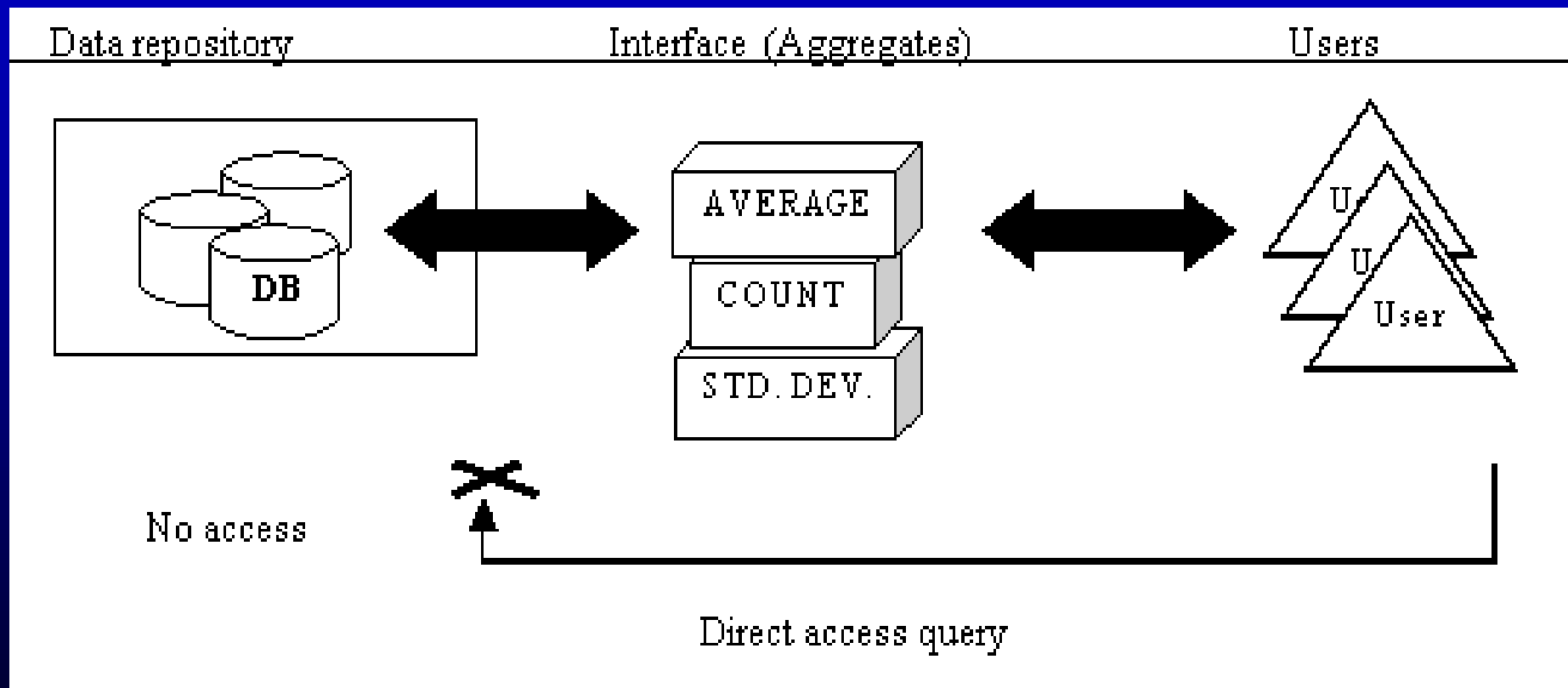


Giới thiệu

- ***CSDL thống kê (SDB)- Statistical database***
 - *Là một CSDL được sử dụng cho mục đích phân tích thống kê.*
 - *Là một CSDL chứa các bản ghi nhạy cảm mô tả về các cá nhân nhưng chỉ các câu truy vấn thống kê (như: COUNT, SUM, AVERAGE, MAX, MIN...) mới được trả lời, ngoài các câu truy vấn này thì những truy vấn vào các mục dữ liệu riêng sẽ không được đáp lại*



Query?



Ví dụ về SDB

- Có 2 dạng SDB cơ bản



Ví dụ về SDB ...

Dạng quan hệ

SDB về công nhân

ID	Tên	Chức vụ	Phòng	Tuổi	Giới tính	Lương
01	Nam	Nhân viên	Marketing	29	M	3500
02	Lan	Trưởng phòng	Kế hoạch	33	F	6200
03	Huệ	Nhân viên	Kế hoạch	27	F	4000
04	Minh	Giám sát viên	Marketing	24	M	3600
05	Quỳnh	Nhân viên	Kế hoạch	24	F	2900

Ví dụ về SDB ...

Dạng quan hệ

SDB về các vụ tai nạn ô tô

HoTen	Tuoi	Đ/C	MauXe	LoaiXe	ThoiGian	CoLoi	SayRuou
Nguyễn Văn Tài	25	HN	Xanh	Honda	13.30	1	1
Lê sỹ Hoàng	37	HD	Đỏ	Toyota	6.25	1	0
Hoàng Văn Minh	42	PT	Trắng	Audi	17.45	0	0
Vũ Bình Minh	32	PT	Vàng	Volkswagon	3.30	0	1
Trần Quang Hòa	22	HN	Xanh	Honda	6.30	1	0

Ví dụ về SDB ...

Dạng quan hệ

SDB về các Sinh viên

Tên	Giới tính	Địa chỉ	Phụ cấp	Lớp
Minh	M	HN	500	Toán1
Hải	M	HD	0	Toán2
Tuyết	F	NĐ	300	Tin1
Nam	M	BG	100	Tin2
Phương	F	NA	200	Toán2
Hạnh	F	HT	100	Toán1

Ví dụ về SDB ...

Dạng quan hệ

SDB về đảng viên

MaDV	HoTen	DiaChi	ChucVu	Luong	DangVien
MA01	Trần Văn Nguyên	Hà Nội	Trưởng phòng	3000	1
MA02	Nguyễn Thị Hoa	Hải Phòng	Nhân viên	2000	0
MA03	Vũ Văn Hiền	Hà Nội	Phó Giám đốc	4000	1
MA04	Trần Thị Mai	Nghệ An	Trưởng phòng	3000	1
MA05	Nguyễn Quang Huy	Hải Phòng	Giám đốc	5000	1
MA06	Trần Văn Hải	Hà Nam	Nhân viên	2000	1
MA07	Lê Minh Sơn	Nam Định	Nhân viên	2500	0

Ví dụ về SDB ...

Dạng vĩ mô

SUM

SDB vĩ mô về các Sinh viên

	AT4A	AT4B	AT4C	AT3
M	500	0	0	100
F	100	200	300	0
Tổng cộng	600	200	300	100

Tổng phụ cấp theo giới tính và theo lớp

Ví dụ về SDB ...

Dạng vĩ mô

COUNT

SDB vĩ mô về công nhân

Năm sinh	Giới tính	Mã phòng		
		Phòng1	Phòng2	Phòng3
1941-1951	M	10	12	0
	F	1	0	3
1952-1962	M	12	10	5
	F	20	2	8
>1962	M	15	0	1
	F	20	10	0

Question

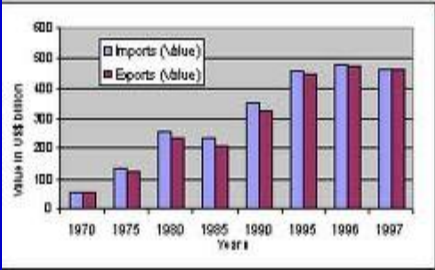
- *Các câu lệnh truy vấn thống kê?*



Ví dụ

NhanVien

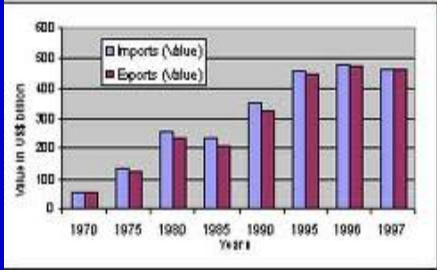
ID	Tên	Chức vụ	Phòng	Tuổi	Giới tính	Lương
01	Nam	Nhân viên	Maketing	29	F	3500
02	Lan	Trưởng phòng	Kế hoạch	33	M	6200
03	Huệ	Nhân viên	Kế hoạch	27	M	3600
04	Minh	Giám sát viên	Maketing	24	F	3600
05	Quỳnh	Nhân viên	Kế hoạch	24	F	2900



Ví dụ một số câu truy vấn thống kê

COUNT

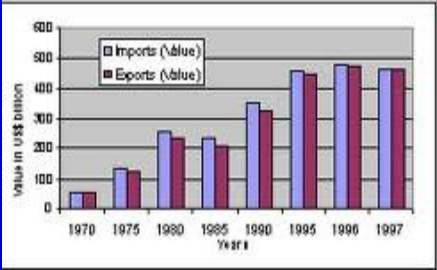
- Select count(*) from Nhanvien
(Trả lại tổng số lượng các bg trong table)
- select count(*) from nhanvien where Luong<=1000
- Select count(Luong) AS count_Luong from Nhanvien
- Select count(Distinct Luong) from



Ví dụ một số câu truy vấn thống kê

SUM

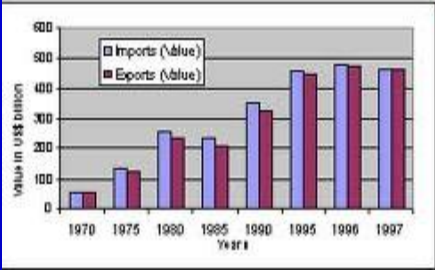
- Select SUM(Luong) as sum_Luong from Nhanvien
- Select SUM(Distinct Luong) as sum_Luong from Nhanvien
- Select Chucvu, Sum(Luong) from Nhanvien GROUP BY chucvu
- Select HoTen, chucvu, Luong from nhanvien ORDER by chucvu
- Compute SUM(Luong) by chucvu



Ví dụ một số câu truy vấn thống kê

AVG

- **Select AVG(Luong) AS avg_Luong from Nhanvien**
- **Select AVG(Luong) AS avg_Luong from Nhanvien where Luong>1000**
- **Select AVG(distinct Luong) AS avg_Luong from Nhanvien**
- **Select chucvu, AVG(Luong) AS avg_Luong, SUM(Luong) as sum_luong from Nhanvien**



Ví dụ một số câu truy vấn thống kê

MIN

- **Select** MIN(Luong) **from** Nhanvien
- **Select** MIN(Distinct Luong) **from** Nhanvien

MAX

- **Select** MAX(Distinct Luong) **from** Nhanvien
- **Select** MAX(Luong) **from** Nhanvien

Question

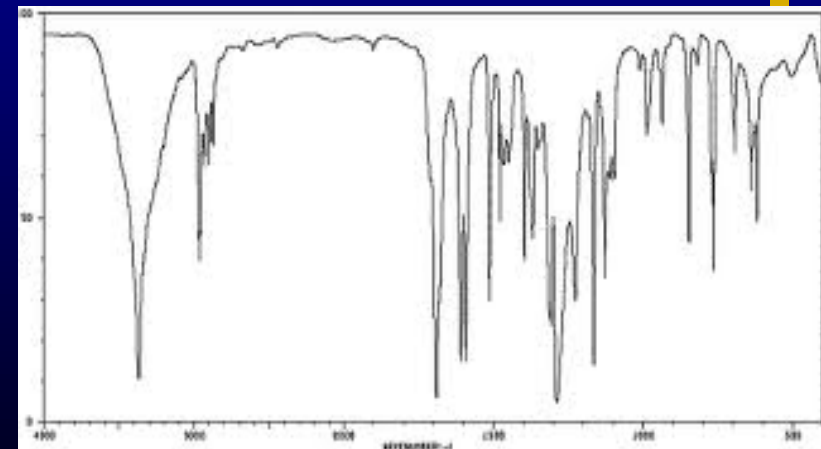
- Ứng dụng và tầm quan trọng của SDB?



What?

Giới thiệu

- **Ứng dụng của SDB (Statistical Database):**
 - Điều tra dân số
 - Thống kê về số người tử vong
 - Về kế hoạch kinh tế
 - Thống kê về khám chữa bệnh
 - Về các vụ tai nạn ô tô
 - Thống kê về tội phạm
 - ...



Ứng dụng của SDB

- Thống kê nông nghiệp, lâm nghiệp, thủy sản
- Thống kê ngành nghề kinh doanh
- Giáo dục và nghiên cứu
- Môi trường
- Thị trường tài chính
- Giá cả và tiêu dùng
- Tài chính công
- Thương mại hàng hoá và dịch vụ



Phân tích và đưa ra chiến lược!

Question

- Ứng dụng SDB ở Việt Nam?
- Ứng dụng SDB trên thế giới?



Question

- **Tại sao phải bảo vệ SDB?**

Dàn xếp giữa:

- Yêu cầu bảo vệ thông tin riêng tư của các cá nhân
- Và quyền truy xuất và xử lý thông tin của các tổ chức

Vì SDB chứa dữ liệu thống kê liên quan đến **thông tin nhạy cảm** của nhiều cá nhân

Question

- Tấn công vào các SDB bằng cách nào?

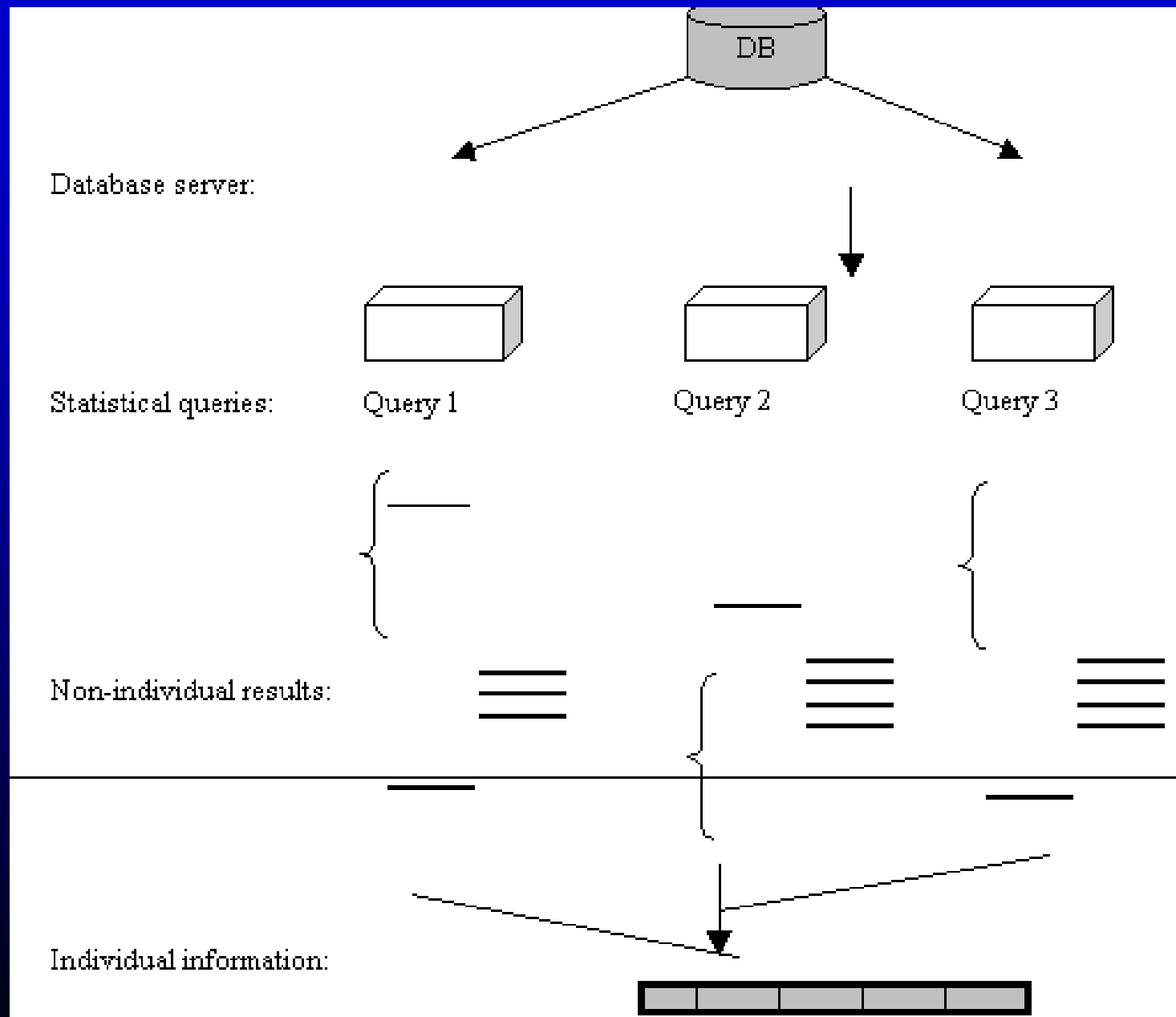
Tấn công suy diễn
(Interference
attack)

Kết hợp các
câu truy vấn
thống kê

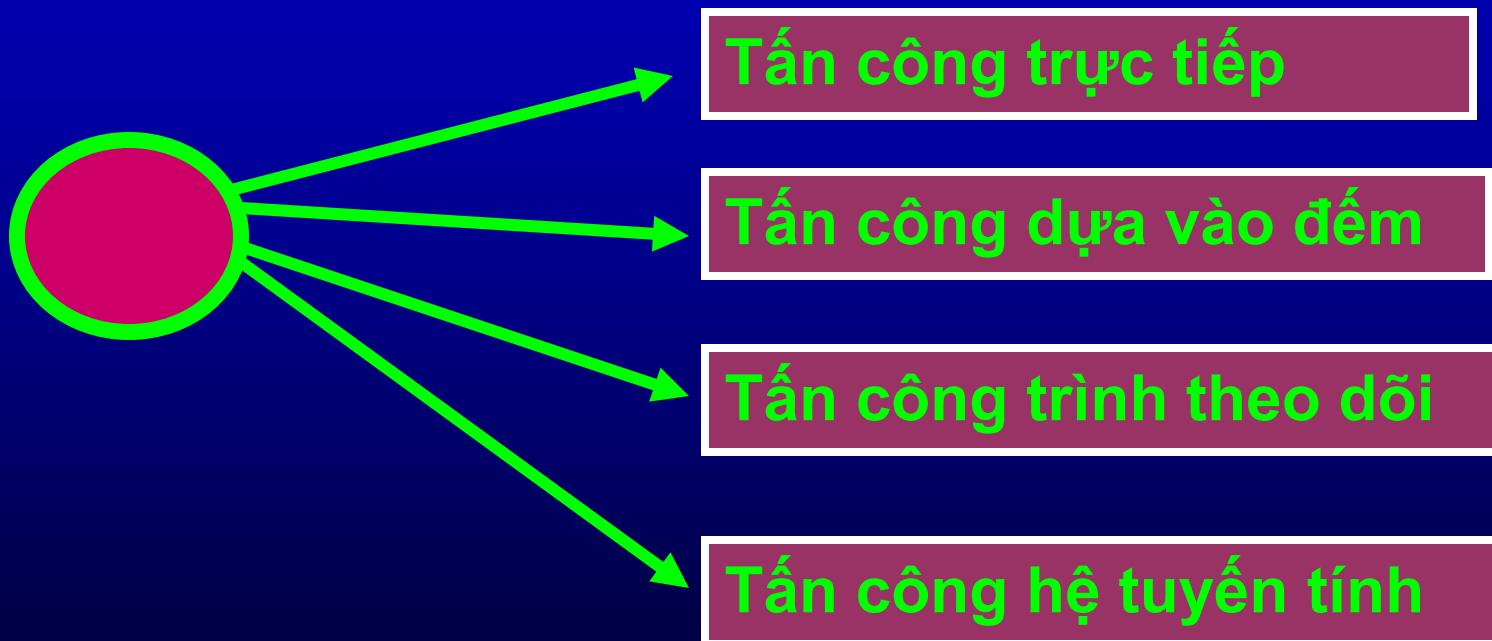
Thu được thông
tin bí mật về 1 cá
nhân

COUNT SUM MIN MAX AVG

Kiến trúc một tấn công suy diễn



Một số kiểu tấn công suy diễn



Nội dung



Giới thiệu



Các khái niệm cơ bản



Một số kiểu tấn công suy diễn



Các kỹ thuật chống tấn công suy diễn

Các khái niệm cơ bản

- ***Các đặc tính của SDB cần được bảo vệ:***

- ***SDB tĩnh:***

- SDB không thay đổi trong suốt thời gian tồn tại của chúng.
 - Ví dụ: CSDL thống kê dân số

- ***SDB động:***

- Thay đổi liên tục theo sự thay đổi của dữ liệu thực, cho phép sửa đổi để phản ánh các thay đổi động của thế giới thực
 - Ví dụ các CSDL nghiên cứu trực tuyến, lớp học trực tuyến khi bổ sung thành viên,....

Các khái niệm cơ bản...

– SDB trực tuyến (online):

- Người sử dụng nhận được các phản hồi thời gian thực cho các câu truy vấn thống kê của mình.

– SDB ngoại tuyến (offline):

- Người sử dụng không biết khi nào các thống kê của họ được xử lý, việc SDB bị lỗi sẽ khó khăn.

Các khái niệm cơ bản...

- *Kiến thức làm việc (working knowledge)*
 - Là tập các mục thông tin (field) và giá trị thuộc tính trong SDB và các kiểu thống kê có sẵn trong SDB mà người dùng có thể biết một cách hợp lệ.
- *Kiến thức bổ sung của người sử dụng (supplementary knowledge):*
 - Người sử dụng có thể có kiến thức bên ngoài về các cá nhân được biểu diễn trong SDB. Người dùng hoàn toàn có thể lợi dụng kiến thức này cho các mục đích xấu để suy diễn.

Mô hình làm lộ SDB

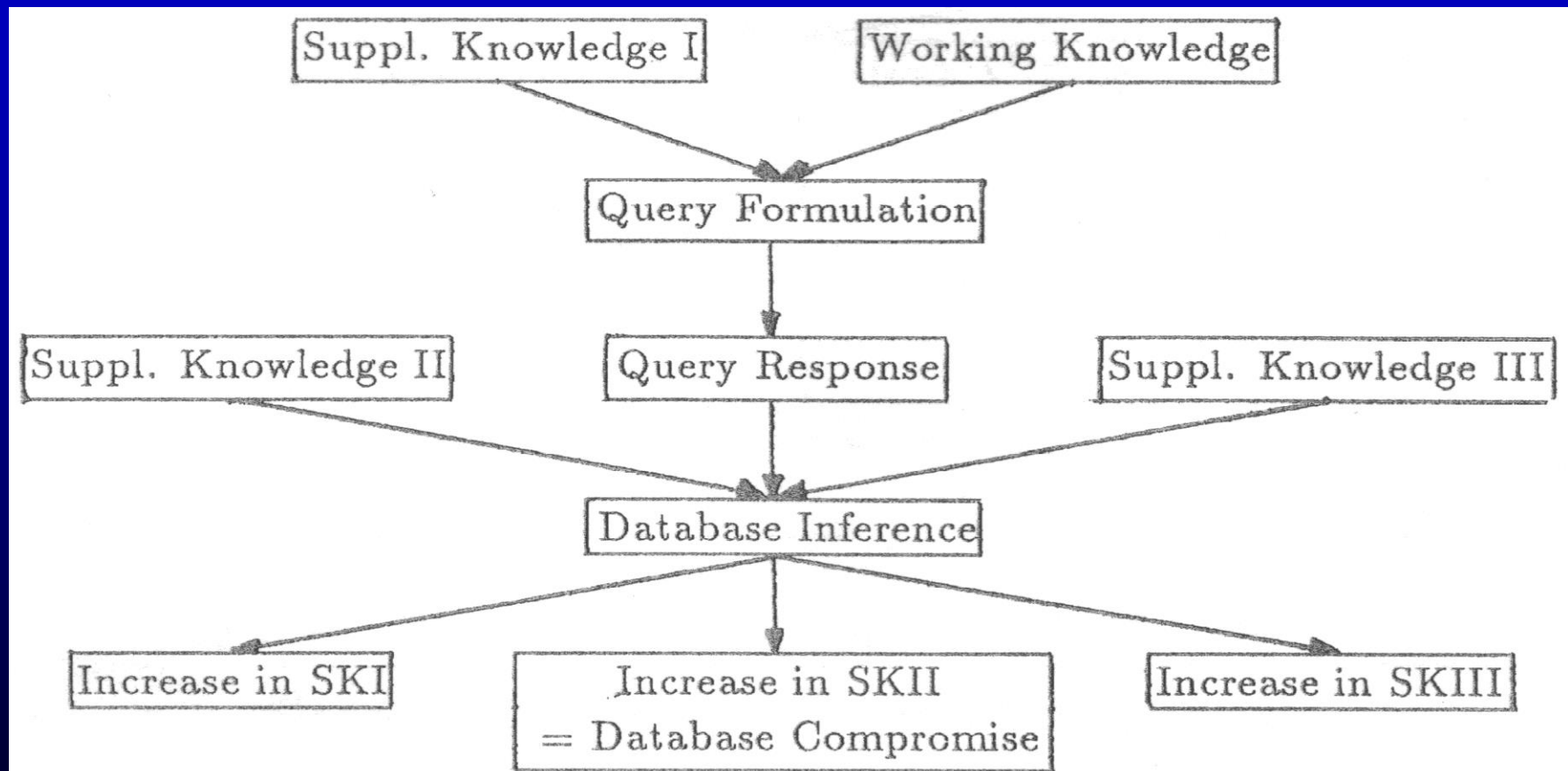


Figure 1. Database compromise and supplementary knowledge.

Ví dụ

NhanVien

ID	Ten	ChucVu	PhongLV	Que	GioiTinh	Luong
01	Nam	Nhân viên	Maketing	Hải Phòng	F	3500
02	Lan	Trưởng phong	Kế hoạch	Hà Nội	M	6200
03	Huệ	Nhân viên	Kế hoạch	Nam Định	M	4000
04	Minh	Giám sát viên	Maketing	Bắc Giang	F	3600
05	Quỳnh	Nhân viên	Kế hoạch	Hà Nội	F	2900

Ví dụ về làm lộ một SDB (Lộ chính xác)

WK = Nam Định \in Que

SK1 = Chỉ có 1 công nhân ở Nam Định

Query { MIN(Luong, que= "Nam Định") }

Result (4000)

Database
Interference

SK2 = Lộ SDB (Lương của 1 công nhân ở Nam Định là 4000)

Ví dụ về làm lộ một SDB (Lộ xấp xỉ)

WK = Giới Tính gồm {M, F}

SK1 = Lương trong khoảng [0, 7000]

Query { MAX(Luong, Giới Tính= "F") }

Result (4500)

Database
Interference

SK2 = Lộ SDB (Lương của tất cả
các nữ công nhân đều ≤ 4500)

Các khái niệm cơ bản...

- *Công thức đặc trưng:*

- Là một công thức lôgíc, được ký hiệu bởi một chữ cái viết hoa (A, B, C, \dots), trong đó các giá trị thuộc tính được kết hợp với nhau thông qua các toán tử Boolean như OR, AND, NOT (\vee, \wedge, \neg).

- Ví dụ:

$$C = (GioiTinh=F) \wedge [(MaPhong="Kế hoạch" \vee (MaPhong="Tài vụ"))] \wedge (NamSinh < 1965)$$

Các khái niệm cơ bản...

- ***Tập truy vấn (query set): của một công thức đặc trưng C là tập tất cả các bản ghi thỏa mãn C.***
 - Ký hiệu là $X(C)$.
- ***Thống kê trên C: là các câu truy vấn thống kê trên C***
 - Ký hiệu: $q(C)$
 - Chẳng hạn: $\text{COUNT}(C)$, $\text{SUM}(C, A_j)$,
 $\text{MIN}(C, A_j)$, $\text{MAX}(C, A_j)$

Ví dụ

NhanVien

ID	Ten	ChucVu	PhongLV	Que	GioiTinh	Lương
01	Nam	Nhân viên	Maketing	Hải Phòng	F	3500
02	Lan	Trưởng phong	Kế hoạch	Hà Nội	M	6200
03	Huệ	Nhân viên	Tài vụ	Nam Định	F	4000
04	Minh	Giám sát viên	Maketing	Bắc Giang	F	3600
05	Quỳnh	Nhân viên	Kế hoạch	Hà Nội	F	2900

Ví dụ

- Công thức đặc trưng C

$$C = \{(GioiTinh = F) \wedge [(MaPhong = \text{"Kế hoạch"}) \vee (MaPhong = \text{"Tài vụ"})]\}$$

- Tập truy vấn X(C)

ID	Ten	ChucVu	PhongLV	Que	GioiTinh	Lương
03	Huệ	Nhân viên	Tài vụ	Nam Định	F	4000
05	Quỳnh	Nhân viên	Kế hoạch	Hà Nội	F	2900

- Thống kê trên C: Count(C), Sum(C, Lương), MAX(C, Lương), MIN(C, Lương)

Các khái niệm cơ bản (...)

- ***Khái niệm bậc:*** Một thống kê gồm m thuộc tính khác nhau được gọi là thống kê bậc m .
 - Ví dụ, $\text{Count}((\text{GioiTinh} = F) \wedge (\text{MaPhong} = \text{Phong1}))$ là một thống kê bậc 2.
 - $\text{Count}(*)$ là thống kê bậc 0.
 - ***Khái niệm thống kê nhạy cảm:*** Thống kê được tính toán trên một *thuộc tính bí mật* trong tập truy vấn có kích cỡ bằng 1 là thống kê nhạy cảm.
 - Ví dụ: $\text{COUNT}(\text{AGE} > 50) = 1$
- => $\text{SUM}(\text{Salary}, \text{age} > 50)$ là thống kê nhạy cảm**

Nội dung



Giới thiệu



Các khái niệm cơ bản

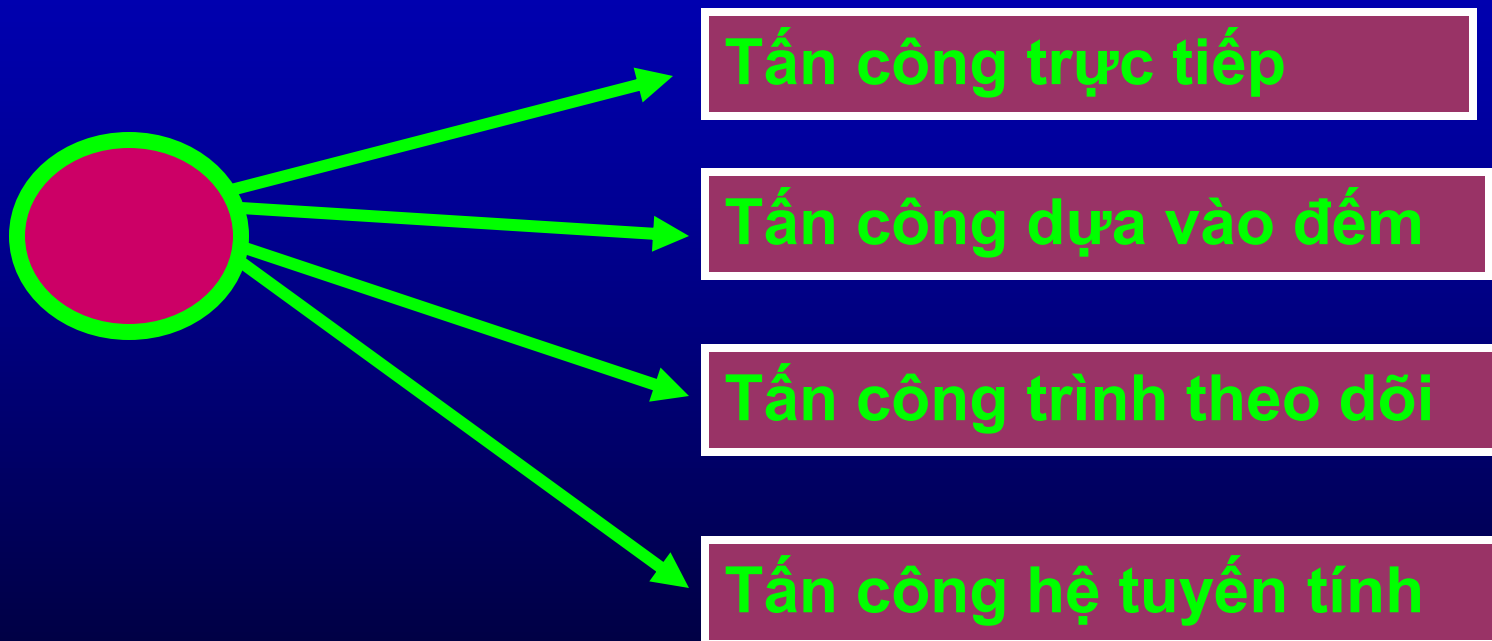


Một số kiểu tấn công suy diễn



Các kỹ thuật chống tấn công suy diễn

Một số kiểu tấn công suy diễn



Một số kiểu tấn công suy diễn...

NhanVien

ID	Tên	Chức vụ	Phòng	Tuổi	Giới tính	Lương
01	Nam	Nhân viên	Maketing	29	F	3500
02	Lan	Trưởng phòng	Kế hoạch	33	M	6200
03	Huệ	Nhân viên	Kế hoạch	27	M	4000
04	Minh	Giám sát viên	Maketing	24	F	3600
05	Quỳnh	Nhân viên	Kế hoạch	24	F	2900

Một số kiểu tấn công suy diễn...

Tấn công trực tiếp:

- Sử dụng các câu truy vấn thông thường, không phải truy vấn thống kê
- Ví dụ:

```
SELECT Ten FROM NhanVien WHERE Luong>4.360
```



Giải pháp

Bộ lọc - Filter (loại các truy vấn không hợp lệ)

Một số kiểu tấn công suy diễn...

Tấn công dựa vào đếm

- Đây là loại tấn công bằng cách kết hợp giá trị đếm với giá trị tổng để thu được thông tin bí mật.
- Ví dụ:

COUNT (ChucVu = “Trưởng phòng”, Phong= “Kế hoạch”) = 1



SUM (Luong, (ChucVu= “Trưởng phòng”, Phong= “Kế hoạch”))

Một số kiểu tấn công suy diễn...

Tấn công trình theo dõi (...)

Tấn công hệ tuyến tính (...)



Sau Kiểm
soát kích cỡ
tập truy vấn

Nội dung



Giới thiệu



Các khái niệm cơ bản



Một số kiểu tấn công suy diễn



Các kỹ thuật chống tấn công suy diễn

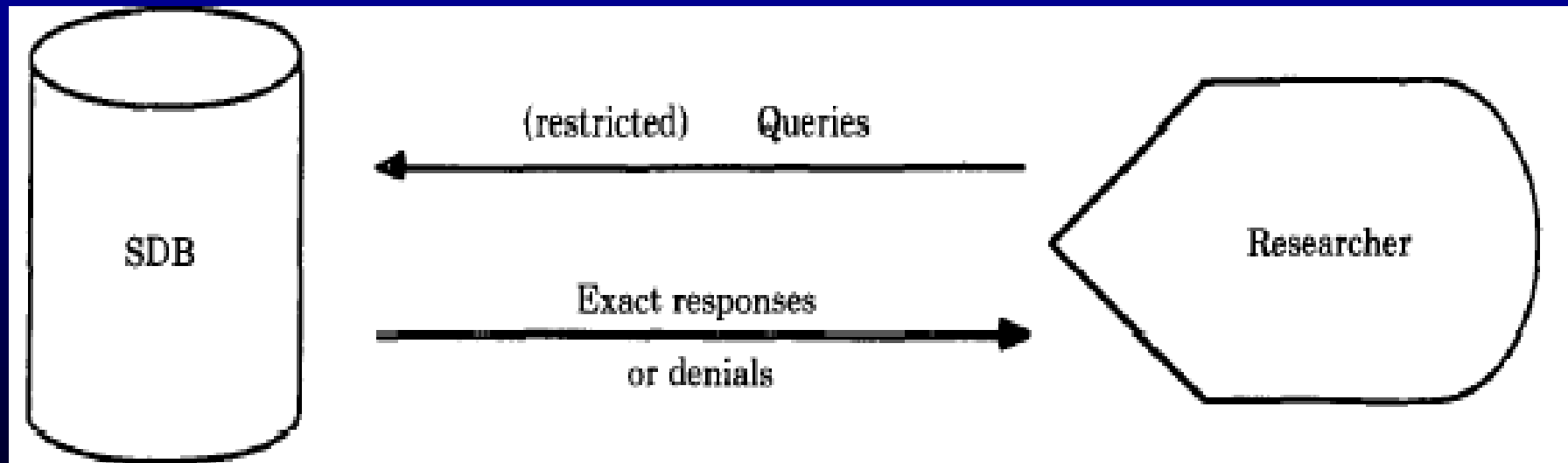
Các kỹ thuật chống suy diễn.

- Từ sự phân loại tổng quát các kỹ thuật chống suy diễn do Denning và Schlorer (1983) và Adam, Wortmann (1989) đưa ra



Tổng quan về các kỹ thuật kiểm soát suy diễn

- **Kỹ thuật khái niệm:** dựa vào mô hình khái niệm
- **Kỹ thuật hạn chế tập truy vấn**



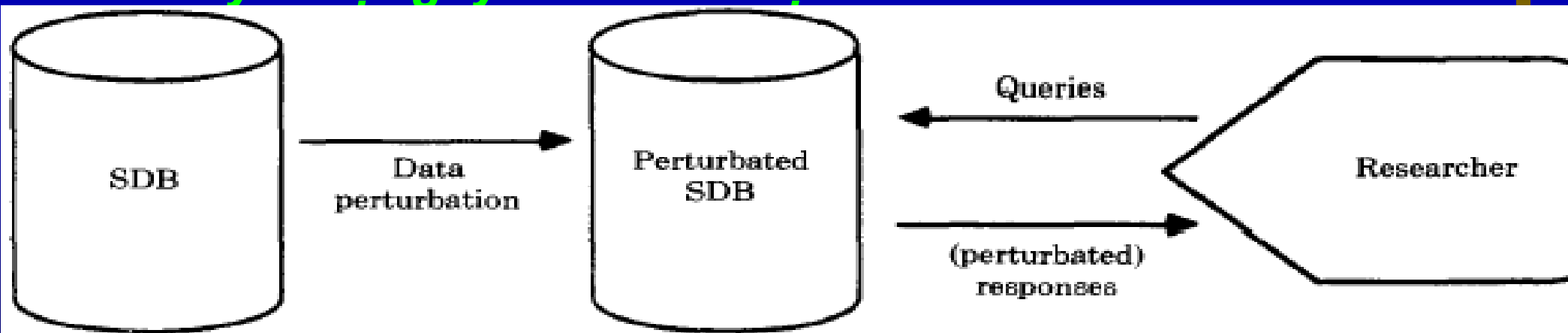
Tổng quan về các kỹ thuật kiểm soát suy diễn

- ***Kỹ thuật dựa vào gây nhiễu***

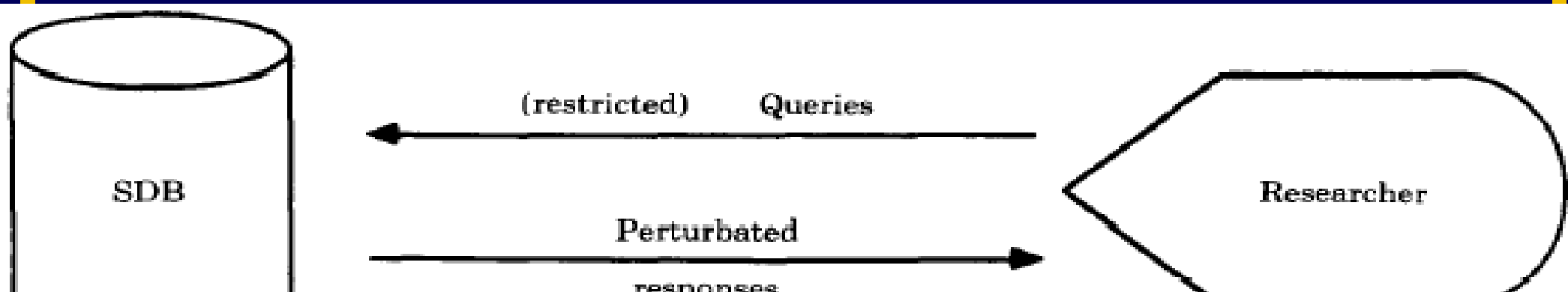
- Làm nhiễu cơ sở dữ liệu thống kê hoặc làm nhiễu kết quả đầu ra của mọi câu truy vấn, bằng cách thêm các “nhiễu”.

Tổng quan về các kỹ thuật kiểm soát suy diễn

- *Kỹ thuật dựa vào gây nhiễu*
 - *Kỹ thuật gây nhiễu dữ liệu*



- *Kỹ thuật gây nhiễu đầu ra*



Các kỹ thuật chống suy diễn.



Kỹ thuật khái niệm

- Làm việc ở mô hình khái niệm của SDB, để tìm ra các tấn công suy diễn có thể có
- Gồm hai kỹ thuật:
 - *Mô hình lưới*
 - *Phân hoạch khái niệm*

Kỹ thuật khái niệm...

- **Mô hình lưới:** do Denning và Schlorer đề xuất, 1983.
 - Là một mô hình khái niệm cung cấp nền tảng cho việc phát hiện những tấn công suy diễn có thể xảy ra với SDB.
 - Xuất phát từ thông tin thống kê được gộp ở nhiều mức khác nhau có thể gây dư thừa dữ liệu
=> người dùng có thể khám phá dữ liệu nhạy cảm.

Kỹ thuật khái niệm...

- **Mô hình lưới:**

- Dựa vào cấu trúc lưới
- Gồm các bảng m-chiều ($0 \leq m \leq N$, N là số thuộc tính của bảng SDB): là các bảng được **gộp** dữ liệu từ một hay nhiều thuộc tính.
- Tính trên một thống kê nào đó như: COUNT, SUM, AVG,...

Ví dụ về SDB ...

Dạng vĩ mô

Ví dụ: mô hình lưới cho SDB về công nhân

COUNT

Bảng 3-chiều (N=3)

Năm sinh	Giới tính	Mã phòng		
		Phòng1	Phòng2	Phòng3
1941-1951	M	10	12	0
	F	1	0	3
1952-1962	M	12	10	5
	F	20	2	8
>1962	M	15	0	1
	F	20	10	0

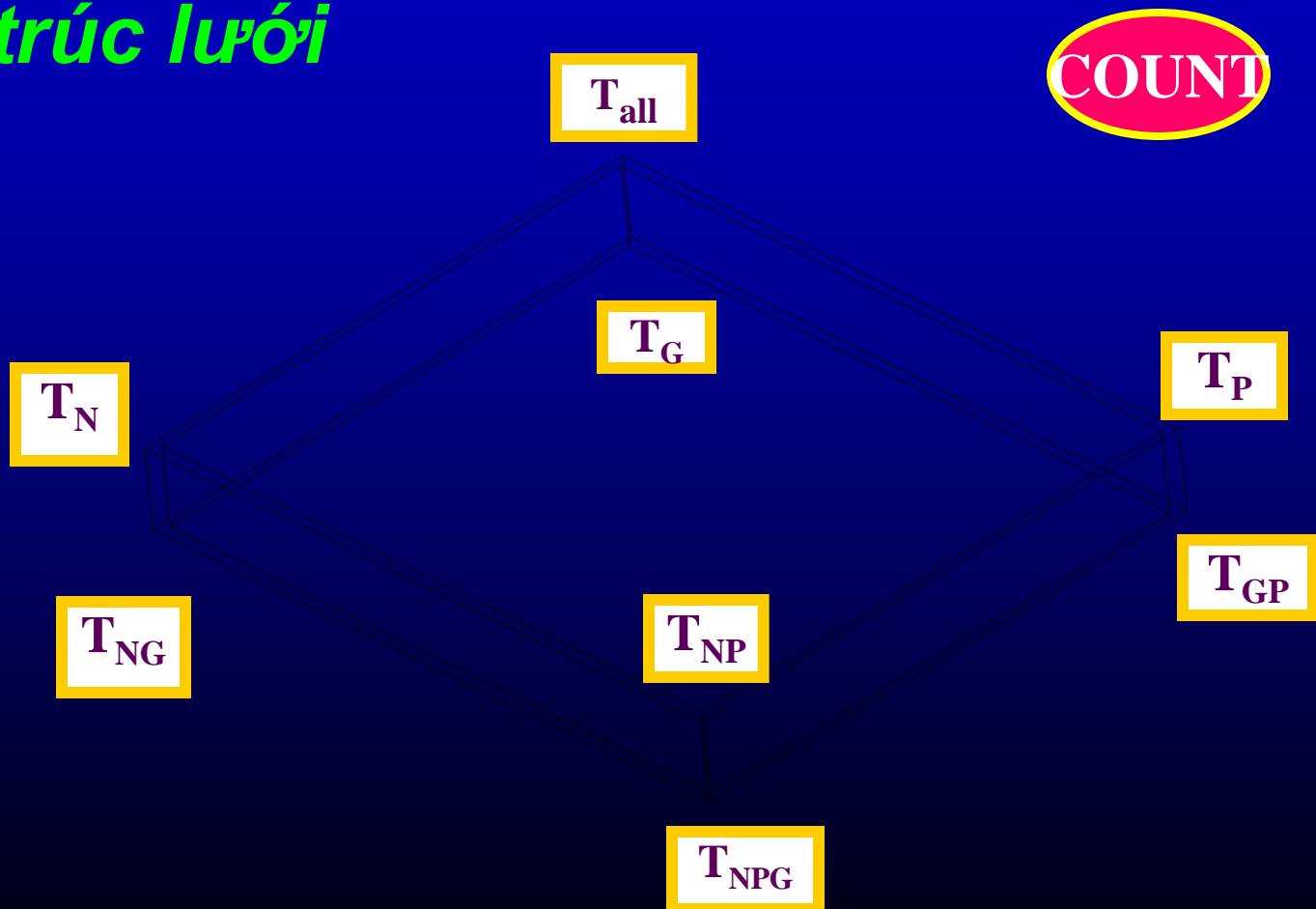
Kỹ thuật khái niệm...

- Cấu trúc lưới*

N: Năm sinh

G: Giới tính

P: Mã phòng



Kỹ thuật khái niệm...

- Các bảng 2-chiều

NG Table		
Năm sinh	Giới tính	
	M	F
1941-1951	22	4
1952-1962	27	30
>1962	16	30

NP Table			
Năm sinh	Mã phòng		
	Phong1	Phong2	Phong3
1941-1951	11	12	3
1952-1962	32	12	13
>1962	35	10	1

SD Table			
Giới tính	Mã phòng		
	Phong1	Phong2	Phong3
M	37	22	6
F	41	12	11

Kỹ thuật khái niệm...

- Các bảng 1-chiều**

Giới tính	
M	F
65	64

Năm sinh	
1941-1951	26
1952-1962	58
>1962	46

Mã phòng		
Phòng1	Phòng2	Phòng3
78	34	17

– **Bảng 0-chiều:**

Kỹ thuật khái niệm...

- **Cấu trúc lưới:**

- **Ưu điểm:** là một mô hình an toàn hiệu quả cho nghiên cứu các vấn đề suy diễn và các phương pháp kiểm soát suy diễn. Với nhiều bảng ở các mức gộp khác nhau, ta có thể phân tích:
 - Các kiểu tấn công suy diễn bằng câu truy vấn COUNT, SUM, AVERAGE,...
 - Các tấn công kiểu kết hợp các câu truy vấn khác nhau để suy diễn ra dữ liệu nhạy cảm...
 - So sánh các kiểm soát suy diễn: hạn chế tập truy vấn và gây nhiễu dữ liệu

Kỹ thuật khái niệm...

- **Cấu trúc lưới:**

- **Nhược điểm:**

- Mô hình lưới không thể cung cấp tính đầy đủ của cơ sở dữ liệu
 - Không phù hợp với cơ sở dữ liệu động, vì khi cập nhật SDB ta phải cập nhật tất cả các bảng trong mô hình lưới, do đó rất tốn công.

Kỹ thuật khái niệm...

- **Phân hoạch khái niệm:** do Chin và Ozsoyoglu đề xuất, 1981.
 - Giải quyết các vấn đề chống suy diễn trong giai đoạn thiết kế khái niệm của SDB.
 - Dựa vào việc định nghĩa tập các cá thể của SDB tại mức khái niệm, được gọi là các **lực lượng (populations)**.
 - Dựa vào các điều kiện cần kiểm tra nhằm tránh suy diễn

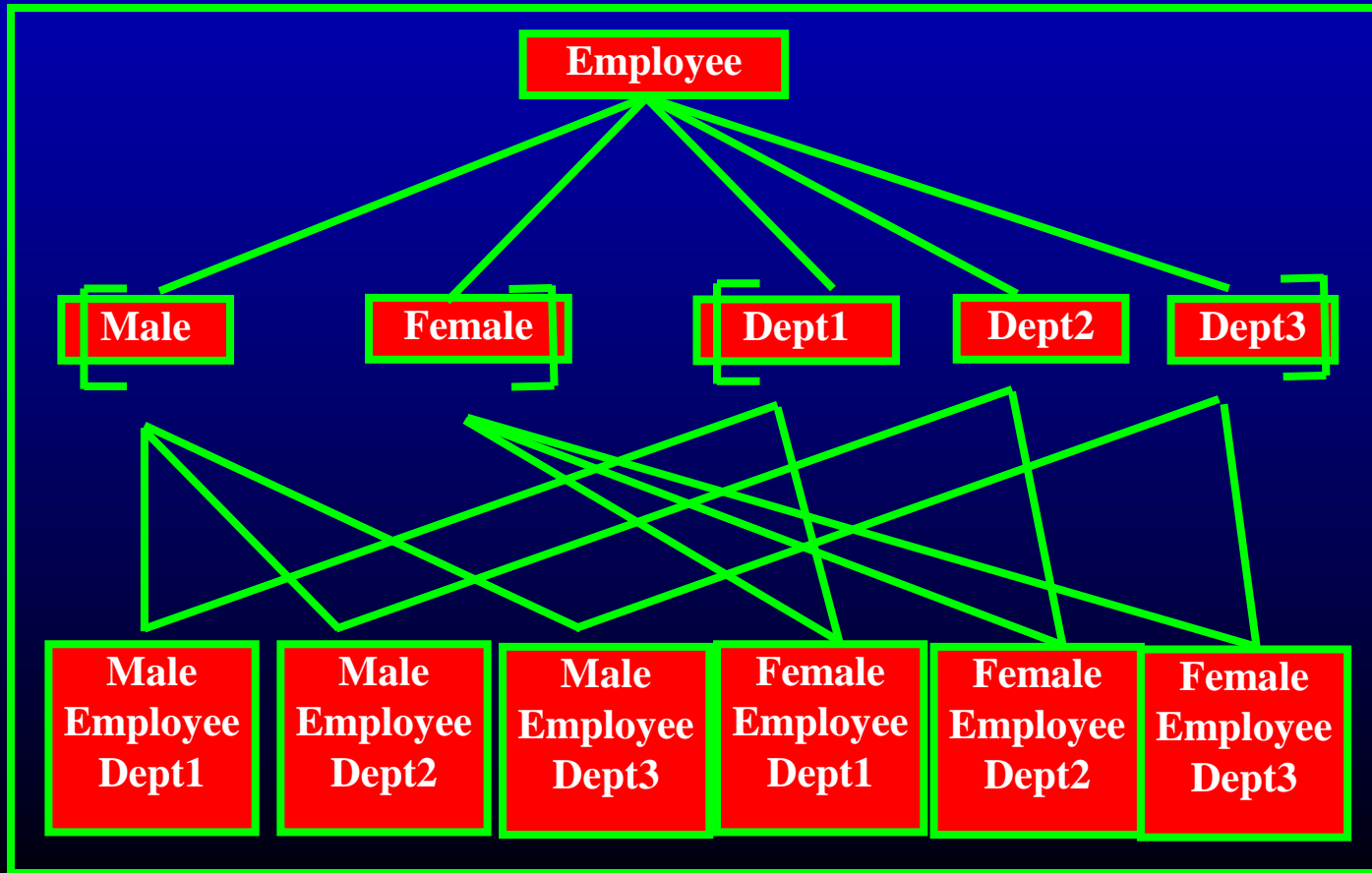
Kỹ thuật khái niệm...

- ***Phân hoạch khái niệm:***

- Hình sau minh họa mô hình khái niệm của một cơ sở dữ liệu thống kê về công nhân - Employee SDB, trong đó lực lượng Employee được phân tách thành ***5 lực lượng con***, tùy thuộc vào các thuộc tính “giới tính” và “Dept-Code”-Mã phòng.
- ***Lực lượng nguyên tử A-Population*** là lực lượng không phân tách được nữa

Kỹ thuật khái niệm...

- *Phân hoạch khái niệm:*



Kỹ thuật khái niệm...

- **Phân hoạch khái niệm:**

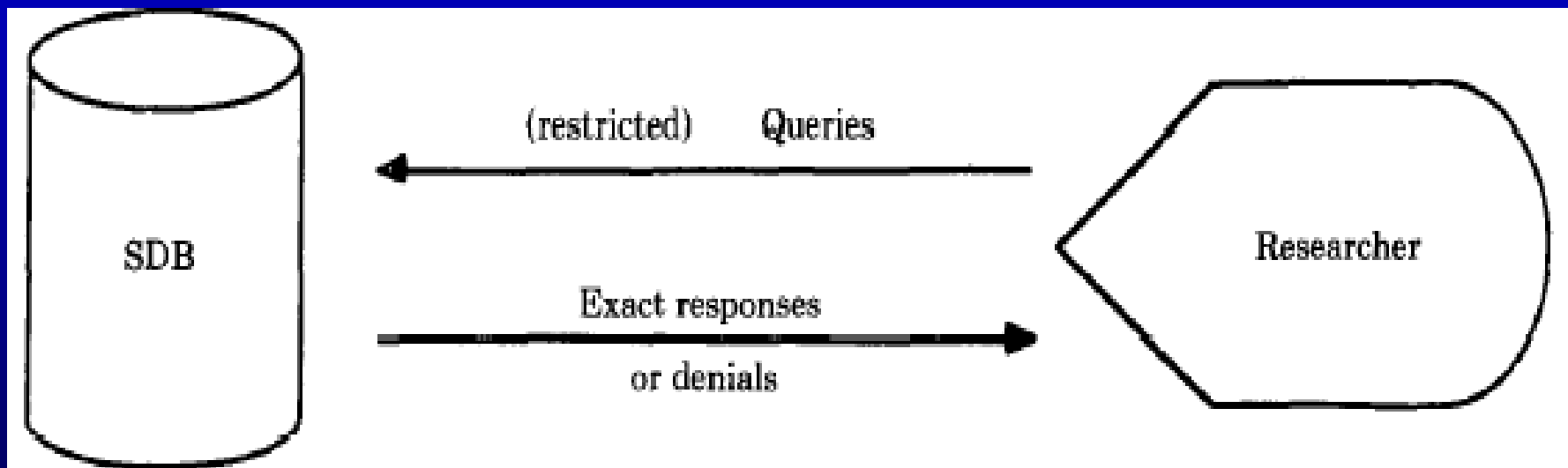
- Để hỗ trợ việc xác định các yêu cầu an toàn thống kê trong mô hình khái niệm này, người ta đã đề xuất hệ thống tiện ích quản lý an toàn thống kê (SSMF) gồm có 3 modul, cụ thể là PDC, UKC và CEC:

- **PDC** (Xây dựng định nghĩa lực lượng- Population Definition Construct)
- **UKC** (Xây dựng trình độ người dùng - User Knowledge Construct)
- **CEC** (Bộ thi hành và kiểm tra ràng buộc - Constraint Enforcer and Checker)

Các kỹ thuật chống suy diễn.



Kỹ thuật hạn chế



- Các kỹ thuật này chống suy diễn bằng cách hạn chế các câu truy vấn thống kê theo một điều kiện hạn chế nào đó

Kỹ thuật hạn chế...

- *Phân loại*

- Kiểm soát kích cỡ tập truy vấn
- Kiểm soát kích cỡ tập truy vấn mở rộng
- Kiểm soát chồng lấp tập truy vấn
- Gộp
- Kỹ thuật giấu ô
- Kỹ thuật kết hợp

Kỹ thuật hạn chế...

- *Kiểm soát kích cỡ tập truy vấn*
- Kiểm soát kích cỡ tập truy vấn mở rộng
- Kiểm soát chồng lấp tập truy vấn
- Gộp
- Kỹ thuật giấu ô
- Kỹ thuật kết hợp

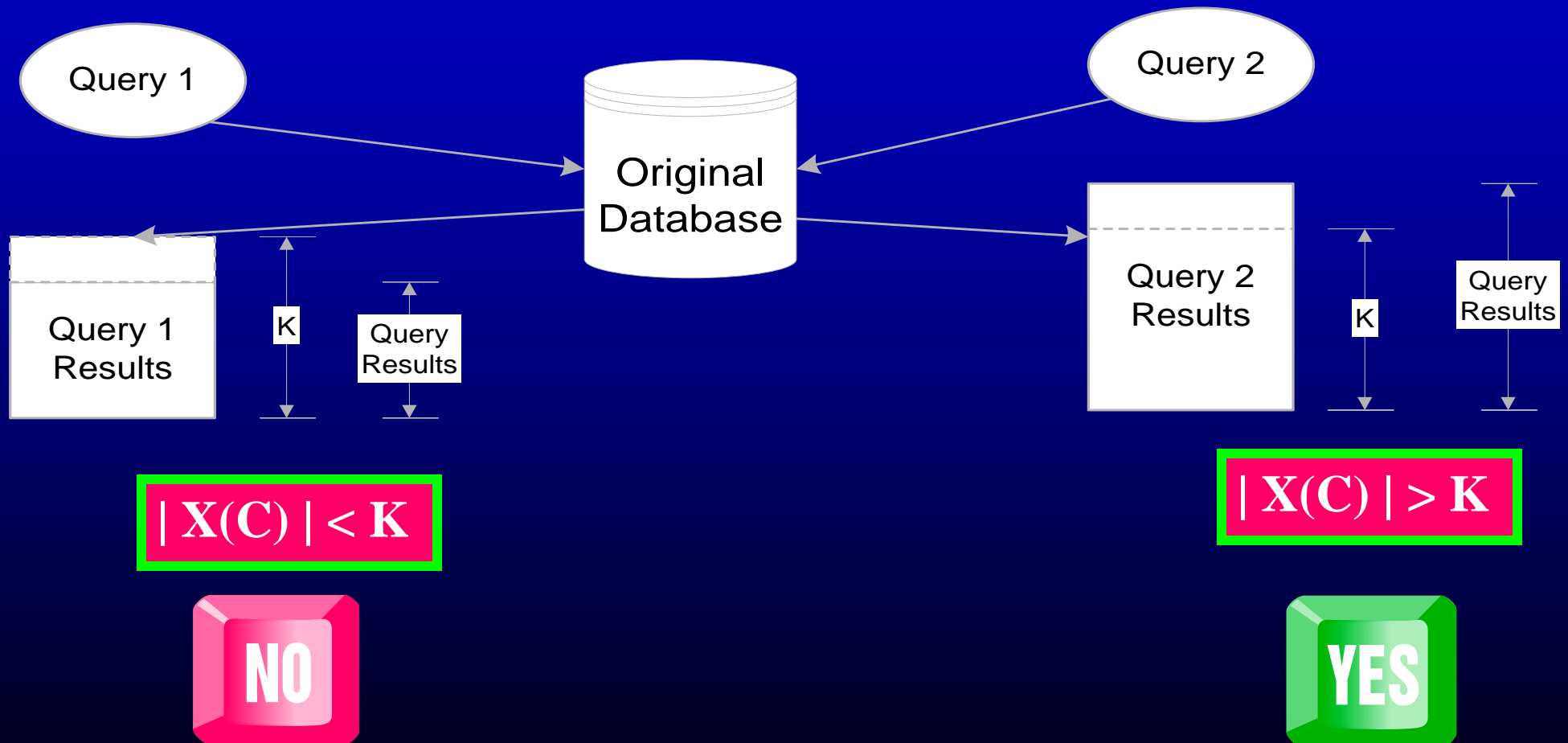
Kiểm soát kích cỡ tập truy vấn

- Một thống kê $q(C)$ chỉ được phép nếu tập truy vấn của nó, $X(C)$, thỏa mãn quan hệ sau:

$$\begin{aligned} K &\leq |X(C)| \leq N-K \\ 0 &\leq K \leq N/2 \end{aligned}$$

- Trong đó, N là tổng số bản ghi trong SDB, K do DBA định nghĩa.

Query Set Restriction



Lưu ý

- **Lưu ý: Trường hợp $K=3$, $K=4$**
 - **Nếu $K=3$** , nghĩa là chặn các truy vấn kích cỡ $=1,2$
 - + Nếu người dùng đoán được $K=3$, và có 1 bản ghi trả về \Rightarrow dễ dàng
 - + Nếu người dùng đoán được $K=3$, và có 2 bản ghi trả về \Rightarrow chỉ cần SD 2 câu truy vấn Min, Max là biết được lương của 2 cá nhân này
 - **Nếu $K=4$** và người dùng đoán được $K=4$
 - + Nếu người dùng biết có 1 hoặc 2 bản ghi trả về \Rightarrow làm tương tự trên
 - + Nếu biết có 3 bản ghi trả về \Rightarrow Chỉ cần SD 3 câu truy vấn (Sum, Min, Max) là tìm ra được cả 3 giá trị lương củ 3 cá nhân.

Kiểm soát kích cỡ tập truy vấn...

- **Ưu điểm:** Kiểm soát này ngăn chặn các tấn công đơn giản, dựa vào các tập truy vấn rất nhỏ hoặc rất lớn.
 - **Ví dụ:**
 - Người dùng yêu cầu thống kê $q_1 = \text{Count}(C) = 1$, \Rightarrow có một cá nhân A thỏa mãn C.
 - Đưa ra thống kê $q_2 = \text{Count}(C \wedge C')$
 - Nếu $q_2 = 1 \Rightarrow A$ thỏa mãn C'
 - Ngược lại, A không thỏa mãn C'
 - Đưa ra thống kê khác, ví dụ $\text{Sum}(C, A_i)$
- \Rightarrow Kiểm soát kích cỡ tập truy vấn không cho phép đưa ra q_1, q_2 .

Chọn $N = 5, K = 2$


$$K \leq |X(C)| \leq N-K$$
$$0 \leq K \leq N/2$$

- Công thức đặc trưng C1

$$C1 = \{(GioiTinh=F) \wedge [(MaPhong="Kế hoạch" \vee (MaPhong="Tài vụ"))]\}$$

- Tập truy vấn $X(C1)$

ID	Ten	ChucVu	PhongLV	GioiTinh	Lương
03	Huệ	Nhân viên	Kế hoạch	F	4000
05	Quỳnh	Nhân viên	Tài vụ	F	2900

$|X(C1)| = 2$  Các thống kê trên C1 được trả lại
 $COUNT(C1), SUM(Lương, C1), AVG(Lương, C1)$

Ví dụ

NhanVien

ID	Ten	ChucVu	Phong	Tuoi	GioiTinh	Luong
01	Nam	Nhân viên	Maketing	29	F	3500
02	Lan	Trưởng phòng	Kế hoạch	33	M	6200
03	Huệ	Nhân viên	Kế hoạch	27	F	4000
04	Minh	Giám sát viên	Maketing	24	F	3600
05	Quỳnh	Nhân viên	Kế hoạch	24	F	2900

Chọn $N = 5, K = 2$


$$K \leq |X(C)| \leq N-K$$
$$0 \leq K \leq N/2$$

- Công thức đặc trưng C2

$C2 = (ChucVu = \text{“Giám sát viên”})$

- Tập truy vấn $X(C2)$

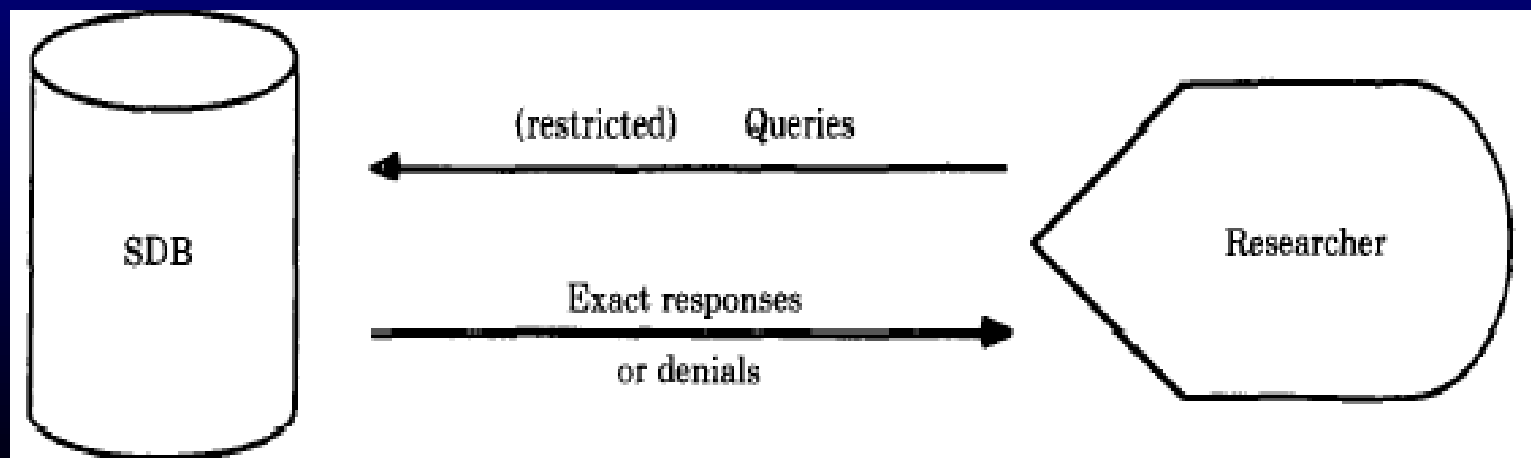
ID	Ten	ChucVu	PhongLV	GioiTinh	Lương
04	Minh	Giám sát viên	Marketing	F	3600

$|X(C2)| = 1$  Các thống kê trên C2 bị chặn
 $COUNT(C2), SUM(Lương, C2), AVG(Lương, C2)$

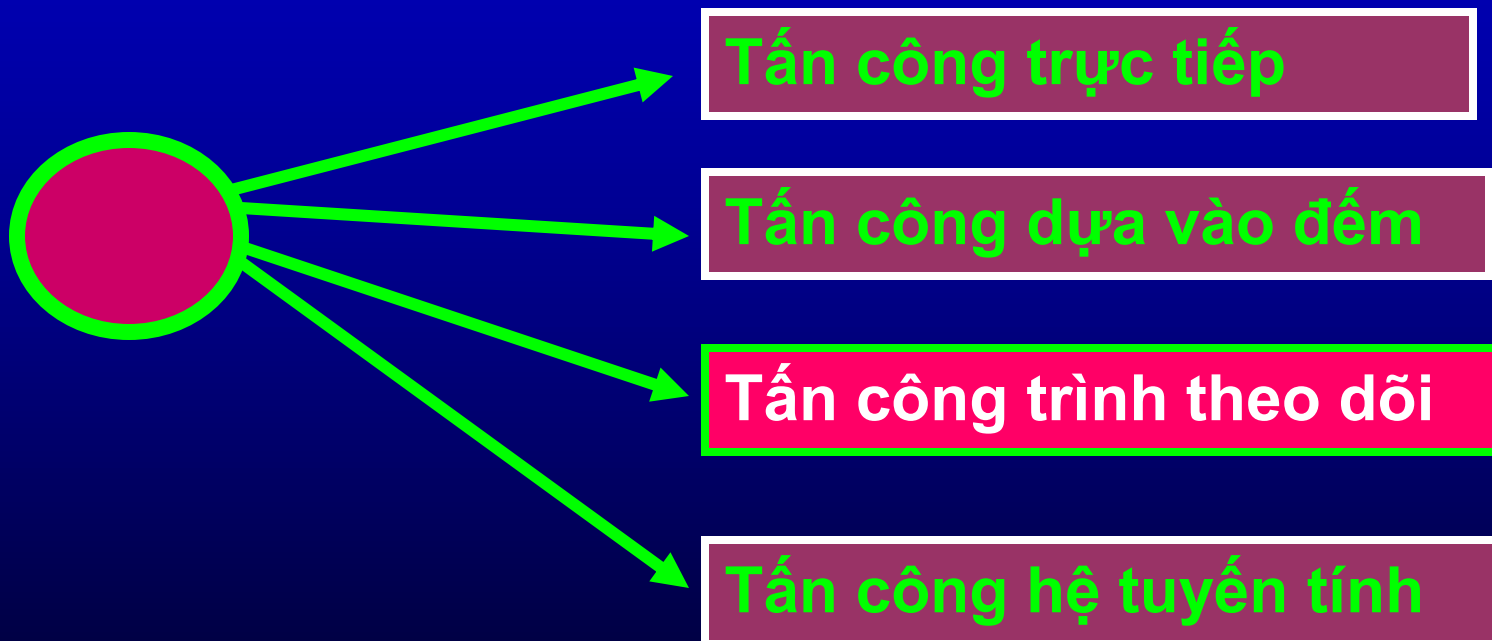
Kiểm soát kích cỡ tập truy vấn...

- *Nhược điểm:*

- Hạn chế khả năng hữu ích của SDB
- Chỉ ngăn chặn được các tấn công đơn giản, khó có thể ngăn chặn được các tấn công phức tạp, như:
Trình theo dõi, Tấn công hệ tuyến tính.



Một số kiểu tấn công suy diễn



Tấn công dựa vào trình theo dõi (Denning&Schlorer)

- *Trình theo dõi (Tracker):*
 - Là một tập các công thức đặc trưng, có thể được sử dụng để đưa thêm bản ghi vào các tập truy vấn kích cỡ nhỏ, làm cho kích cỡ của chúng nằm trong khoảng $[k, N-k]$.
 - Thông qua các trình theo dõi có thể tính toán được các thống kê bị hạn chế.

Tấn công trình theo dõi...

- Giả sử **C** là công thức đặc trưng người dùng yêu cầu
- **T** là một trình theo dõi. T thỏa mãn điều kiện:

$$K \leq |X(T)| \leq N-K$$

Tấn công trình theo dõi...

Kiểu 1

- *Giả thiết:*

K=2

- User cần tính $\text{Count}(C)$, $\text{Sum}(C, \text{Luong})$
- Công thức $C = (A \wedge B)$, và $\text{Count}(C) = 1$.

➡ *Câu truy vấn này bị cấm!*

- *Tấn công:*

- Tính $T = A \wedge \neg B$ thỏa mãn $k \leq |X(T)| \leq N-k$.
- Tính gián tiếp $\text{Count}(C)$:

$$Q(C) = Q(A \wedge B) = Q(A) - Q(A \wedge \neg B)$$

$$\Rightarrow Q(C) = Q(A) - Q(T)$$

Ví dụ

NhanVien

ID	Ten	ChucVu	Phong	Tuoi	GioiTinh	Luong
01	Nam	Nhân viên	Maketing	29	F	3500
02	Lan	Trưởng phòng	Kế hoạch	33	M	6200
03	Huệ	Nhân viên	Kế hoạch	27	F	4000
04	Minh	Giám sát viên	Maketing	24	F	3600
05	Quỳnh	Nhân viên	Kế hoạch	24	F	2900

Ví dụ

Giả thiết

$$C = (\text{Phong} = \text{'Kế hoạch'}) \wedge (\text{Tuoi} = 24) \wedge (\text{GioiTinh} = F)$$

- User cần tính $\text{Count}(C)$
- $\text{Count}(C) = 1$. \Rightarrow Câu truy vấn này bị cấm!

N=5

K=2

• Tấn công:

+ Đặt $C = (A \wedge B)$

$A = (\text{Phong} = \text{'Kế hoạch'})$

$B = (\text{Tuoi} = 24) \wedge (\text{GioiTinh} = F)$

+ Tính $\text{Count}(T) = \text{Count}(A \wedge \neg B) = 2$ thỏa mãn
 $2 \leq \text{Count}(T) = 2 \leq 3$.

+ Tính gián tiếp $\text{Count}(C)$:

$\text{Count}(C) = \text{Count}(A \wedge B)$

$= \text{Count}(A) - \text{Count}(A \wedge \neg B)$

$\text{Count}(C) = \text{Count}(A) - \text{Count}(T) = 3 - 2 = 1$

Ví dụ...

- Tấn công....:

+ Đặt $C = (A \wedge B)$

$A = (\text{Phong} = \text{'Kế hoạch'})$

$B = (\text{Tuoi} = 24) \wedge (\text{Gioi Tinh} = F)$

+ Tính gián tiếp $\text{Sum}(C, \text{Luong})$:

$\text{Sum}(C, \text{Luong}) = \text{Sum}(A \wedge B, \text{Luong})$

$= \text{Sum}(A, \text{Luong}) - \text{Sum}(A \wedge \neg B, \text{Luong})$

$\text{Sum}(C, \text{Luong}) = (6200 + 4000 + 2900) - (6200 + 4000)$
 $= 2900$

→ Đây chính là lương của nhân viên Quỳnh

Bài tập 1

NhanVien

ID	Ten	ChucVu	Phong	Tuoi	GioiTinh	Luong
01	Nam	Nhân viên	Marketing	29	F	3500
02	Lan	Trưởng phòng	Kế hoạch	33	M	6200
03	Huệ	Nhân viên	Kế hoạch	27	F	4000
04	Minh	Giám sát viên	Marketing	24	F	3600
05	Nam	Nhân viên	Kế hoạch	24	M	2900
06	Yến	Nhân viên	Tài vụ	40	F	4600
07	Nam	Phó phòng	Tài vụ	38	M	5000

$C = (Ten = \text{“Nam”}) \wedge (ChucVu = \text{“Phó phòng”})$

Tấn công trình theo dõi...

Kiểu 2 *Giả thiết:*

- Cần tính $\text{Count}(C)$, $\text{Count}(C) < k$
Thống kê này bị cấm!



• *Tấn công:*

- Chọn T thỏa mãn: $k \leq |X(T)|$, $|X(\neg T)| \leq N - k$.
- $Q(D) = Q(\text{All}) = Q(T) + Q(\neg T)$ ($Q(\text{All})$ bị cấm)
- Tính gián tiếp $Q(C)$:

$$Q(C) = Q(C \vee T) + Q(C \vee T^c) - Q(D)$$

Ví dụ SDB về các vụ tai nạn mô tô

HoTen	Tuoi	Đ/C	MauXe	LoaiXe	ThoiGian	CoLoi	SayRuo u
Tài	25	HN	Xanh	Honda	13.30	1	1
Hoàng	37	HD	Đỏ	Toyota	6.25	1	0
Minh	42	PT	Trắng	Honda	17.45	1	0
Minh	19	PT	Vàng	Volkswagon	3.30	0	1
Hòa	22	HN	Xanh	Honda	6.30	1	0

Ví dụ SDB về các vụ tai nạn mô tô

- ***Giả thiết: $C = (Ten='Minh') \wedge (MauXe='Trắng')$***

– ***Count(C)=1, SUM(CoLoi, C)=1***

2 Câu truy vấn này bị cấm!

N=5

K=2

 ***Tấn công:***

– ***Chọn $T = (Tuoi < 25) \Rightarrow Count(T)=2, Count(\neg T)=3$***

– ***Count(All)= Count(T) + Count($\neg T$) = 5***

– ***Tính: $Count(C) = Count(C \vee T) + Count(C \vee \neg T) - Count(All)$***

$$= 3 + 3 - 5 = 1$$

– ***SUM(CoLoi, C)= Sum(CoLoi, $C \vee Tuoi < 25$) + Sum(CoLoi, $C \vee Tuoi \geq 25$) – Sum(CoLoi, All)***



$$= 2 + 3 - 4 = 1.$$

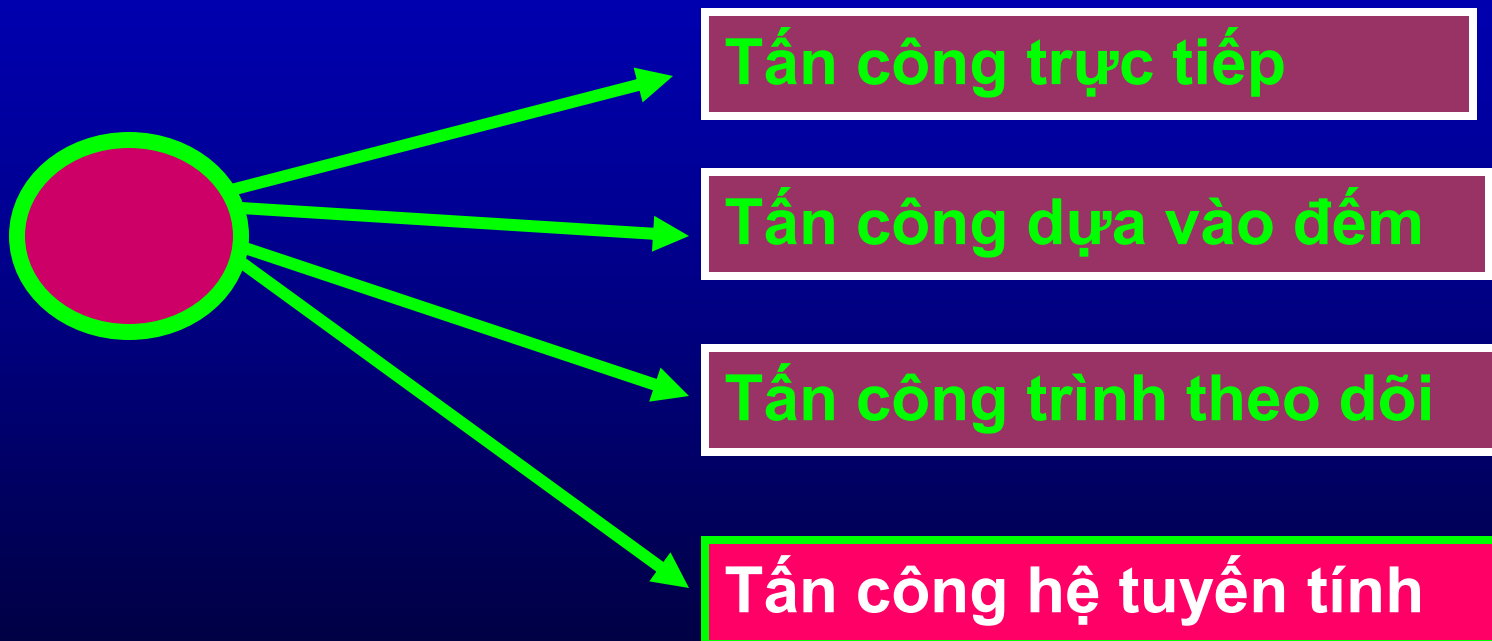
Bài tập 2

NhanVien

ID	Ten	ChucVu	Phong	Tuoi	GioiTinh	Luong
01	Nam	Nhân viên	Maketing	29	F	3500
02	Lan	Trưởng phòng	Kế hoạch	33	M	6200
03	Huệ	Nhân viên	Kế hoạch	27	F	4000
04	Minh	Giám sát viên	Maketing	24	F	3600
05	Nam	Nhân viên	Kế hoạch	24	M	2900
06	Yến	Nhân viên	Tài vụ	40	F	4600
07	Nam	Phó phòng	Tài vụ	38	M	5000

$C = (Ten = \text{“Nam”}) \wedge (ChucVu = \text{“Phó phòng”})$

Một số kiểu tấn công suy diễn



Tấn công hệ tuyến tính

- Là loại tấn công bằng cách giải một hệ phương trình có dạng: $HX = Q$

$$\lambda_{1,1}x_1 + \lambda_{1,2}x_2 + \dots + \lambda_{1,n}x_N = q_1$$

$$\lambda_{2,1}x_1 + \lambda_{2,2}x_2 + \dots + \lambda_{2,N}x_N = q_2$$

.

.

$$\lambda_{k,1}x_1 + \lambda_{k,2}x_2 + \dots + \lambda_{k,n}x_N = q_K$$

Mỗi phương trình tương ứng một câu truy vấn

Tấn công hệ tuyến tính...

- H là ma trận truy vấn
 - $H[i,j] = 1$ nếu bản ghi $x_j \in X(C_i)$, (tương ứng q_i)
 - $H[i,j] = 0$ nếu ngược lại

$$H = \begin{vmatrix} \lambda_{1,1} & \lambda_{1,2} & \cdot & \cdot & \cdot & \lambda_{1,n} \\ \lambda_{2,1} & \lambda_{2,2} & \cdot & \cdot & \cdot & \lambda_{2,n} \\ \vdots & \vdots & \dots & & & \vdots \\ \lambda_{k,1} & \lambda_{k,2} & \cdot & \cdot & \cdot & \lambda_{k,n} \end{vmatrix}$$

- x_1, \dots, x_N là giá trị của N bản ghi
- $Q = (q_1, \dots, q_k)$ là vector của các thống kê đưa ra

Ví dụ

NhanVien

ID	Tên	Chức vụ	Phòng	Tuổi	Giới tính	Lương
01	Nam	Nhân viên	Tai vụ	29	F	3500
02	Lan	Trưởng phòng	Kế hoạch	33	M	6200
03	Huệ	Nhân viên	Kế hoạch	27	M	4000
04	Minh	Giám sát viên	Marketing	24	F	3600
05	Quỳnh	Nhân viên	Kế hoạch	24	F	2900

Tấn công hệ tuyến tính...

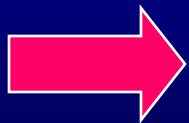
Giả thiết

- $C = (Phong = 'Kế hoạch') \wedge (GioiTinh = F)$
- Cần tính $q = \text{Count}(C) = 1 \rightarrow$ Bị chặn!

Thực hiện

Tính $q_1 = \text{Count}(Phong = 'Kế hoạch')$

- Tính $q_2 = \text{Count}(Phong = 'Kế hoạch', GioiTinh = M)$



$$\begin{cases} q_1 = 0x_1 + 1x_2 + 1x_3 + 0x_4 + 1x_5 = 3 \\ q_2 = 0x_1 + 1x_2 + 1x_3 + 0x_4 + 0x_5 = 2 \end{cases}$$

$$\begin{aligned} \Rightarrow q_3 &= \text{Count}(Phong = 'Kế hoạch', GioiTinh = F) \\ &= q_1 - q_2 = 3 - 2 = 1. \end{aligned}$$

$$\Rightarrow \mathbf{q = q_3 = 1}$$

Ví dụ SDB về công nhân:

- $C = (Phong='Kế\ hoạch') \wedge (GioiTinh=F)$
 - Cần tính $q = Sum(Luong, C)$
 - Tính $q_1 = X(C_1) = Count(Phong='Kế\ hoạch') = 3$
 - Tính $q_2 = X(C_2) = Count(Phong='Kế\ hoạch', GioiTinh=M)=2$
 - $Sum(Luong, C) = Sum(Luong, C_1) - Sum(Luong, C_2)$
 $= (6200+4000+2900) - (6200+4000) = 2900.$
 - Như vậy, kẻ tấn công đã tìm ra lương của người thỏa mãn C.

Tấn công hệ tuyến tính:

- Ví dụ

- Giả sử cần tính $q_3 = \text{Sum}(\text{Sex} = M \wedge \text{Dept-Code} = \text{Dept3} \wedge \text{Birth-Year} = 1968, \text{Salary})$, $\text{count} = 1$.

$$\begin{cases} q_1 = \text{Sum}(\text{Sex} = F \wedge \text{Dept-Code} = \text{Dept3} \wedge \text{Birth-Year} = 1968, \text{Salary}) \\ q_2 = \text{Sum}((\text{Sex} = F \vee \text{Sex} = M) \wedge \text{Dept-Code} = \text{Dept3} \wedge \text{Birth-Year} = 1968, \text{Salary}) \end{cases}$$

- Tương ứng ta có hệ sau:
- $\text{Count1} = 7 \quad \begin{cases} x_1 + x_3 + x_4 + x_6 + x_7 + x_8 + x_9 = 33 \end{cases}$
- $\text{Count2} = 8 \quad \begin{cases} x_1 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 = 37 \end{cases}$
- Từ đó, tính được: $x_5 = q_2 - q_1 = 4$.
- Và người dùng biết $\text{count}(q_3) = 1 \Rightarrow$ tìm được lương của người này

Kiểm soát kích cỡ tập truy vấn...

- **Ưu điểm:**

- Đưa ra kết quả chính xác
- Chỉ chống được tấn công suy diễn đơn giản

- **Nhược điểm:**

- Không chống được một số tấn công phức tạp như: Trình theo dõi, Hệ tuyến tính.
- Hạn chế khả năng hữu ích của SDB (vì hạn chế nhiều câu truy vấn)

Kỹ thuật dựa vào hạn chế

- Kiểm soát kích cỡ tập truy vấn
- *Kiểm soát kích cỡ tập truy vấn mở rộng*
- Kiểm soát chồng lấp tập truy vấn
- Kiểm soát dựa vào kiểm toán
- Gộp
- Kỹ thuật giấu ô
- Kỹ thuật kết hợp

Kiểm soát kích cỡ tập truy vấn mở rộng

- ★ • Nhược điểm của kiểm soát kích cỡ tập truy vấn là do các công thức đặc trưng liên quan đến nhau (ví dụ: C và T).
- ➔ • **Cải tiến:** tăng số lượng các tập truy vấn cần được kiểm soát.

Cho công thức đặc trưng C

- + *Tìm tập truy vấn ngầm định của C*
- + *Kiểm soát kích cỡ tập truy vấn với cả tập này*

Kiểm soát kích cỡ tập truy vấn mở rộng

- Cho trước một thống kê bậc m có dạng như sau:

$$q(A1 = a1 \wedge A2 = a2 \wedge \dots \wedge Am = am) \text{ Hoặc } q(A1 = a1 \vee A2 = a2 \vee \dots \vee Am = am)$$

- Khi đó, tồn tại $2^m = C_m^0 + C_m^1 + C_m^2 + \dots + C_m^{m-1}$ tập truy vấn ngầm định, tương ứng với các thống kê sau đây:

$$q(A1 = a1 \wedge A2 = a2 \wedge \dots \wedge Am = am)$$

$$q(A1 = a1 \wedge A2 = a2 \wedge \dots \wedge \neg Am = am)$$

.....

$$q(A1 = a1 \wedge \neg A2 = a2 \wedge \dots \wedge Am = am)$$

$$q(\neg A1 = a1 \wedge A2 = a2 \wedge \dots \wedge Am = am)$$

...

$$q(\neg A1 = a1 \vee \neg A2 = a2 \vee \dots \vee \neg Am = am)$$

Kiểm soát kích cỡ tập truy vấn mở rộng



• Ưu điểm:

- Chống được các kiểu tấn công: Trình theo dõi, Hệ tuyến tính



• Nhược điểm:

- Phải kiểm tra 2^m tập truy vấn ngầm định (hàm mũ tăng rất lớn theo m) => Rất tốn công!



Giải pháp này khó thực hiện

- Ngoài tập truy vấn ngầm định, kẻ tấn công có thể sử dụng những công thức khác liên quan đến tập truy vấn này để tính ra truy vấn yêu cầu.

Ví dụ: tấn công ngoài tập truy vấn ngầm định

- Chúng ta xét 2 thuộc tính A_i và A_j trong SDB
- A_i có n giá trị (a_{i1}, \dots, a_{in}) và A_j có p giá trị (a_{j1}, \dots, a_{jp})
- Xét câu truy vấn tổng kê:

$$q(A_i \wedge A_j)$$

➔ Tạo thành $n \times p$ câu truy vấn con:

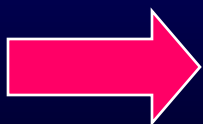
- $q(A_i=a_{i1} \wedge A_j=a_{j1}), \dots, q(A_i=a_{i1} \wedge A_j=a_{jp})$
- $q(A_i=a_{i2} \wedge A_j=a_{j1}), \dots, q(A_i=a_{i2} \wedge A_j=a_{jp})$
- ...
- $q(A_i=a_{in} \wedge A_j=a_{j1}), \dots, q(A_i=a_{in} \wedge A_j=a_{jp})$

Ví dụ: tấn công ngoài tập truy vấn ngầm định

- Trong các câu truy vấn trên, giả thiết chỉ có truy vấn sau là nhạy cảm:

$$q(A_i=a_{i1} \wedge A_j=a_{j1}) = q(a_{i1} \wedge a_{j1})$$

- Tập truy vấn ngầm định gồm: $2^2 = 4$ tập truy vấn:
 - $q(a_{i1} \wedge a_{j1})$, $q(a_{i1} \wedge \neg a_{j1})$
 - $q(\neg a_{i1} \wedge a_{j1})$, $q(\neg a_{i1} \wedge \neg a_{j1})$.



4 câu truy vấn này sẽ bị cấm theo KS kích cỡ tập truy vấn mở rộng!

Ví dụ: tấn công ngoài tập truy vấn ngầm định

- Tuy nhiên, kẻ tấn công có thể thực hiện như sau:

$$q(a_{i_1} \wedge a_{j_1}) = q(a_{j_1}) - \frac{q(a_{j_1} \wedge \neg a_{i_1})}{(Bị\ cấm)}$$

$$= q(a_{j_1}) - \frac{[q(a_{j_1} \wedge a_{i_2}) + \dots + q(a_{j_1} \wedge a_{i_n})]}{(Không\ bị\ cấm)}$$

Kỹ thuật dựa vào hạn chế...

- Kiểm soát kích cỡ tập truy vấn
- Kiểm soát kích cỡ tập truy vấn mở rộng
- *Kiểm soát chồng lấp tập truy vấn*
- Gộp
- Kỹ thuật giấu ô
- Kỹ thuật kết hợp

Kỹ thuật dựa vào hạn chế...

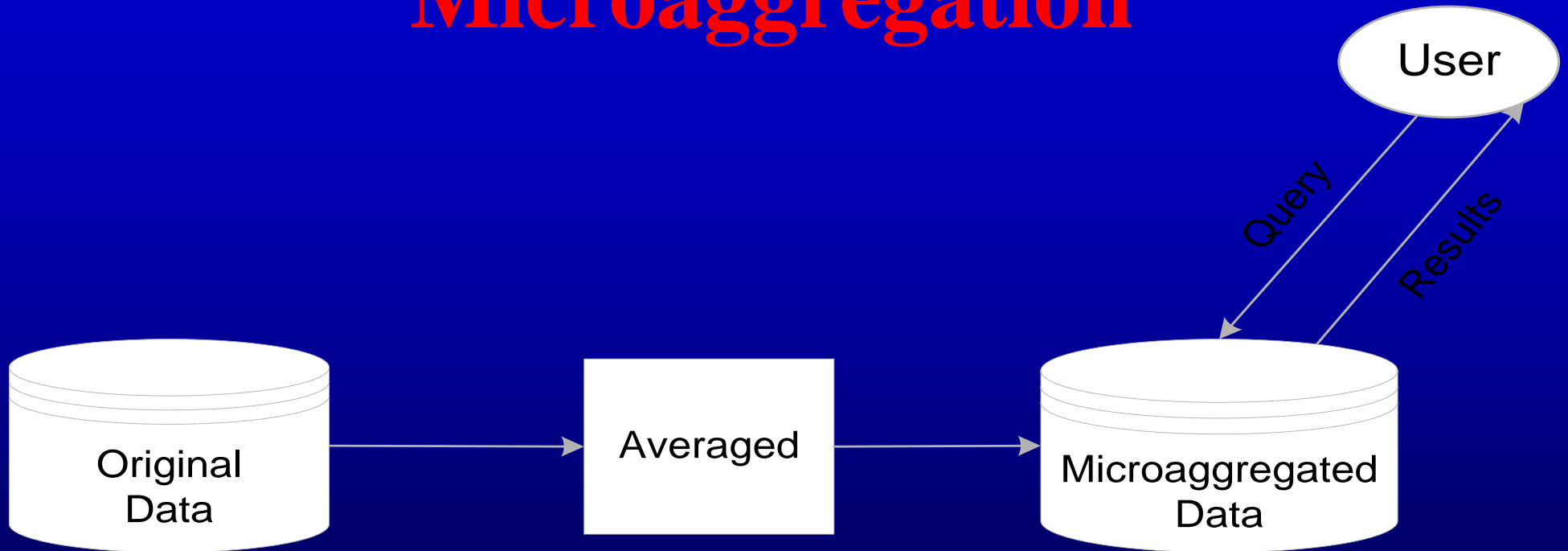
- Kiểm soát kích cỡ tập truy vấn
- Kiểm soát kích cỡ tập truy vấn mở rộng
- Kiểm soát chồng lấp tập truy vấn
- **Gộp**
- Kỹ thuật giấu ô
- Kỹ thuật kết hợp

Kỹ thuật gộp (*microaggregation*)



- Các câu truy vấn thống kê được tính toán trên các **nhóm gộp**. Dữ liệu riêng sẽ được nhóm lại thành một khối nhỏ trước khi đưa ra.
- **Giá trị trung bình** của **nhóm gộp** sẽ thay thế cho mỗi giá trị riêng của dữ liệu được gộp
- Kỹ thuật này giúp ngăn chặn khám phá dữ liệu riêng.

Microaggregation

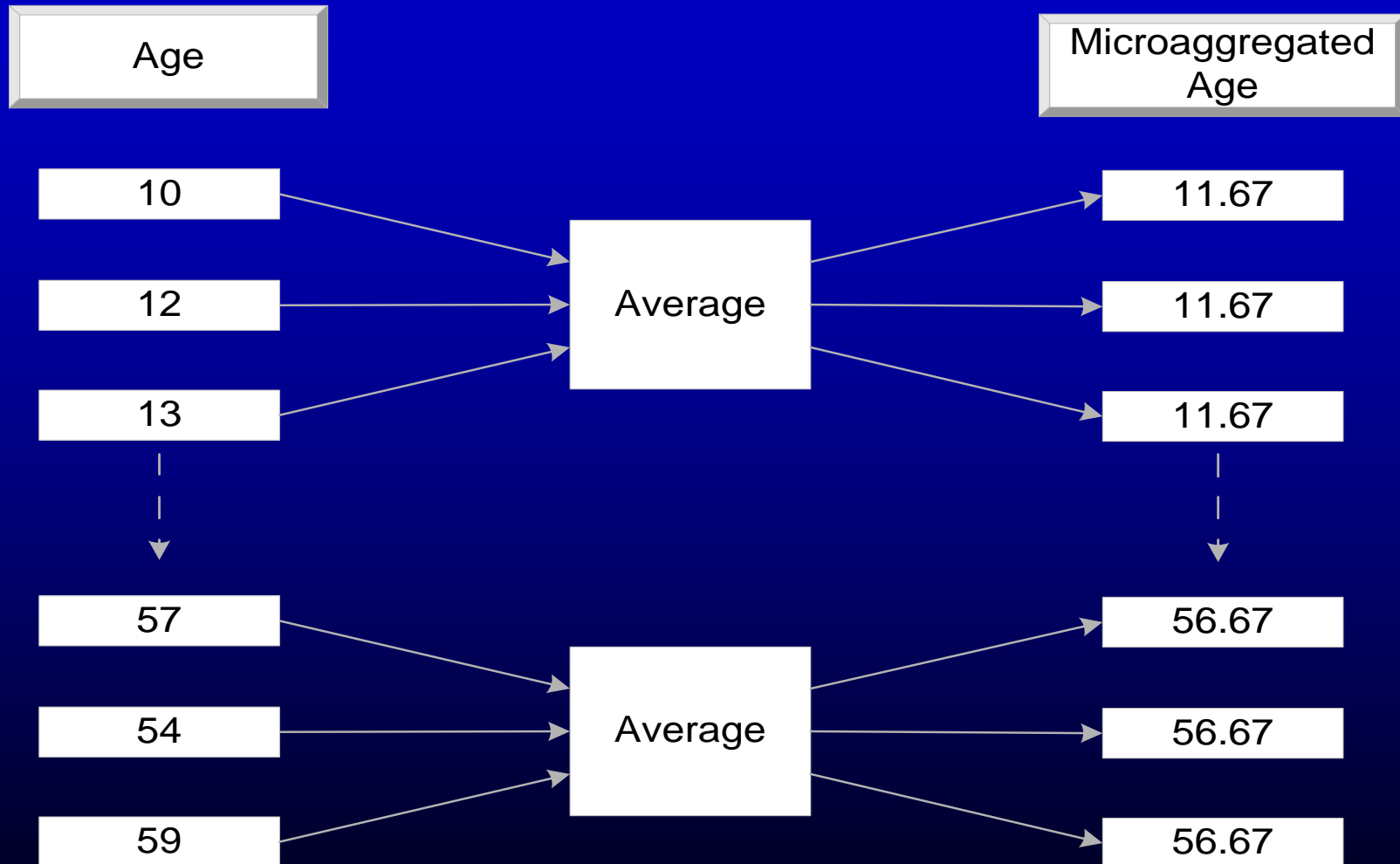


Kỹ thuật gộp (*microaggregation*)...

- Ví dụ:

- Cục thống kê nông nghiệp quốc gia (NASS) công bố dữ liệu về các nông trường, trang trại.
- Để bảo vệ chống lại sự khám phá dữ liệu, dữ liệu chỉ được đưa ra ở **mức vùng**.
- Dữ liệu tại các nông trại ở mỗi vùng sẽ được gộp để bảo vệ tính riêng tư và tránh bị khám phá.

Microaggregation



Kỹ thuật gộp (*microaggregation*)...



- **Ưu điểm:**

- Tránh được việc để lộ thông tin nhạy cảm

- **Nhược điểm:**

- Kết quả đưa ra không chính xác

Kỹ thuật dựa vào hạn chế...

- Kiểm soát kích cỡ tập truy vấn
- Kiểm soát kích cỡ tập truy vấn mở rộng
- Kiểm soát chồng lấp tập truy vấn
- Gộp
- *Kỹ thuật giấu ô*
- Kỹ thuật kết hợp

Kỹ thuật Giấu ô (Cell suppression)

- Kỹ thuật này được thiết kế cho các *SDB vĩ mô* (đưa ra các thống kê trong bảng 2- chiều, như các thống kê dân số).

- ★ • ***Giấu ô***: trong các bảng
 - Giấu đi tất cả các ô tương ứng với các thống kê nhạy cảm
 - Giấu thêm các ô tương ứng với các thống kê có thể gián tiếp khám phá ra các thống kê nhạy cảm (*Giấu bổ sung*).

Kỹ thuật Giấu ô (Cell suppression)

- ***Tiêu chuẩn giấu ô:***

- ***Thống kê Count:*** kích cỡ tập truy vấn nhỏ hơn hoặc bằng 1, nghĩa là $\text{Count}(C) = 0$, $\text{Count}(C) = 1$
- ***Thống kê Sum,*** tiêu chuẩn nhạy cảm được sử dụng là quy tắc «*đáp ứng n , trội $k\%$* » .
 - “*Nếu tổng n hoặc ít hơn n bản ghi giá trị một thuộc tính tạo thành $k\%$ hoặc lớn hơn $k\%$ trong toàn bộ thống kê Sum của ô đó*” \Rightarrow ô này bị giấu
 - Các tham số n và k được giữ bí mật và do DBA xác định ($n < N$)

ID	Ten	ChucVu	Phong	Tuoi	GioiTinh	Luong
01	Nam	Nhân viên	Maketing	29	F	3500
02	Lan	Trưởng phòng	Maketing	33	F	6200
03	Huệ	Nhân viên	Kế hoạch	27	M	4000
04	Minh	Giám sát viên	Maketing	24	F	3600
05	Bình	Nhân viên	Tài vụ	23	F	2000
06	Hải	Nhân viên	Kế hoạch	25	M	1500
07	Hiền	Nhân viên	Tài vụ	21	F	1700
08	Thành	Nhân viên	Kế hoạch	20	M	3000
09	Trường	Phó phòng	Kế hoạch	27	M	5000
10	Bích	Nhân viên	Tài vụ	33	F	600
11	Hoàng	Phó phòng	Kế hoạch	35	M	2500
12	Phượng	Nhân viên	Kế hoạch	52	F	4500
13	Cường	Trưởng phòng	Tài vụ	34	F	6900
14	Việt	Nhân viên	Marketing	57	F	5000
15	Minh	Nhân viên	Tài vụ	37	M	600

Ví dụ

- Từ CSDL trên, ta có CSDL thống kê tổng lương của các công nhân theo Phòng và theo độ tuổi.

**n=1,
k=90%**

**n=2,
k=90%**

?

Tuổi	Phòng			Tổng lương
	Kế hoạch	Maketing	Tài vụ	
<27	4500 ₍₂₎	3600 ₍₁₎	3700 ₍₂₎	11800
27-30	9000 ₍₂₎	3500 ₍₁₎	0 ₍₀₎	12500
>30	7000 ₍₂₎	11200 ₍₂₎	8100 ₍₃₎	27200
Tổng lương	20500	18300	12700	51500

Ví dụ

- Ví dụ: Giả sử $n = 2$ và $k = 90\%$

SUM

Địa chỉ	Mã phòng			Tổng lương
	Phòng1	Phòng2	Phòng3	
Hà Nội	135	80	50	265
Hải Phòng	120	360	100	580
Nam Định	225	90	900	1215
Nghệ An	300	210	75	585
Tổng lương	780	740	1125	2645

Tổng phụ cấp của nam,nữ công nhân trong các phòng

Kỹ thuật Giấu ô (Cell suppression)

Ví dụ 1

- Giả sử kết quả giấu ô như sau:

Địa chỉ	Mã phòng			Tổng lương
	Phòng1	Phòng2	Phòng3	
Hà Nội	135	80	50	265
Hải Phòng	120	360	----	580
Nam Định	----	90	900	1215
Nghệ An	300	210	75	585
Tổng lương	780	740	1125	2645

Kỹ thuật Giấu ô (Cell suppression)



Cần giấu ô bổ sung?

Kỹ thuật Giấu ô (Cell suppression)

Ví dụ 2

- Giả sử kết quả giấu ô như sau:

Địa chỉ	Mã phòng			Tổng lượng
	Phòng1	Phòng2	Phòng3	
Hà Nội	135	80	50	265
Hải Phòng	120	360	----	580
Nam Định	225	90	900	1215
Nghệ An	300	210	75	585
Tổng lượng	780	740	1125	2645

Kỹ thuật Giấu ô (Cell suppression)

 Cần giấu ô bổ sung

Địa chỉ	Mã phòng			Tổng lượng
	Phòng1	Phòng2	Phòng3	
Hà Nội	135	80	50	265
Hải Phòng	-----	360	-----	580
Nam Định	225	90	900	1215
Nghệ An	-----	210	-----	585
Tổng lượng	780	740	1125	2645

Kỹ thuật Giấu ô (Cell suppression)

- *Ưu điểm:*

- Chống được các tấn công kết hợp dựa vào Count và Sum

- *Nhược điểm:*

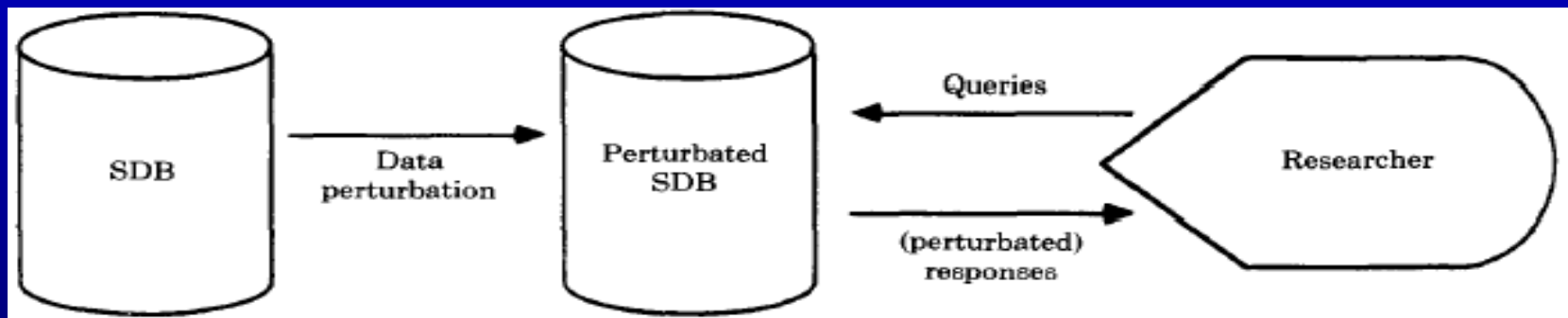
- Hạn chế khả năng hữu ích của SDB, vì phải che giấu một số ô trong CSDL.

Các kỹ thuật chống suy diễn.

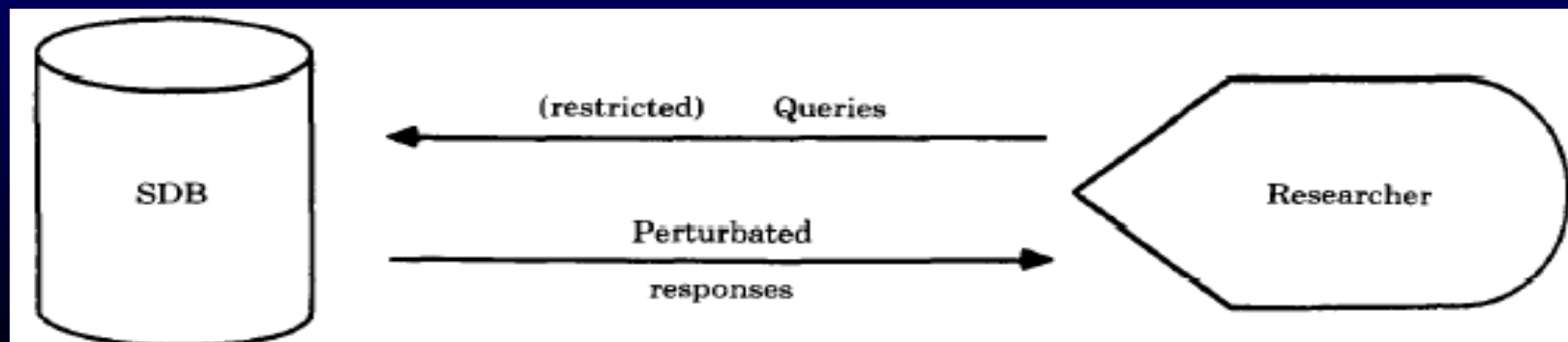


Các kỹ thuật dựa vào gây nhiễu

– Kỹ thuật gây nhiễu dữ liệu



– Kỹ thuật gây nhiễu đầu ra



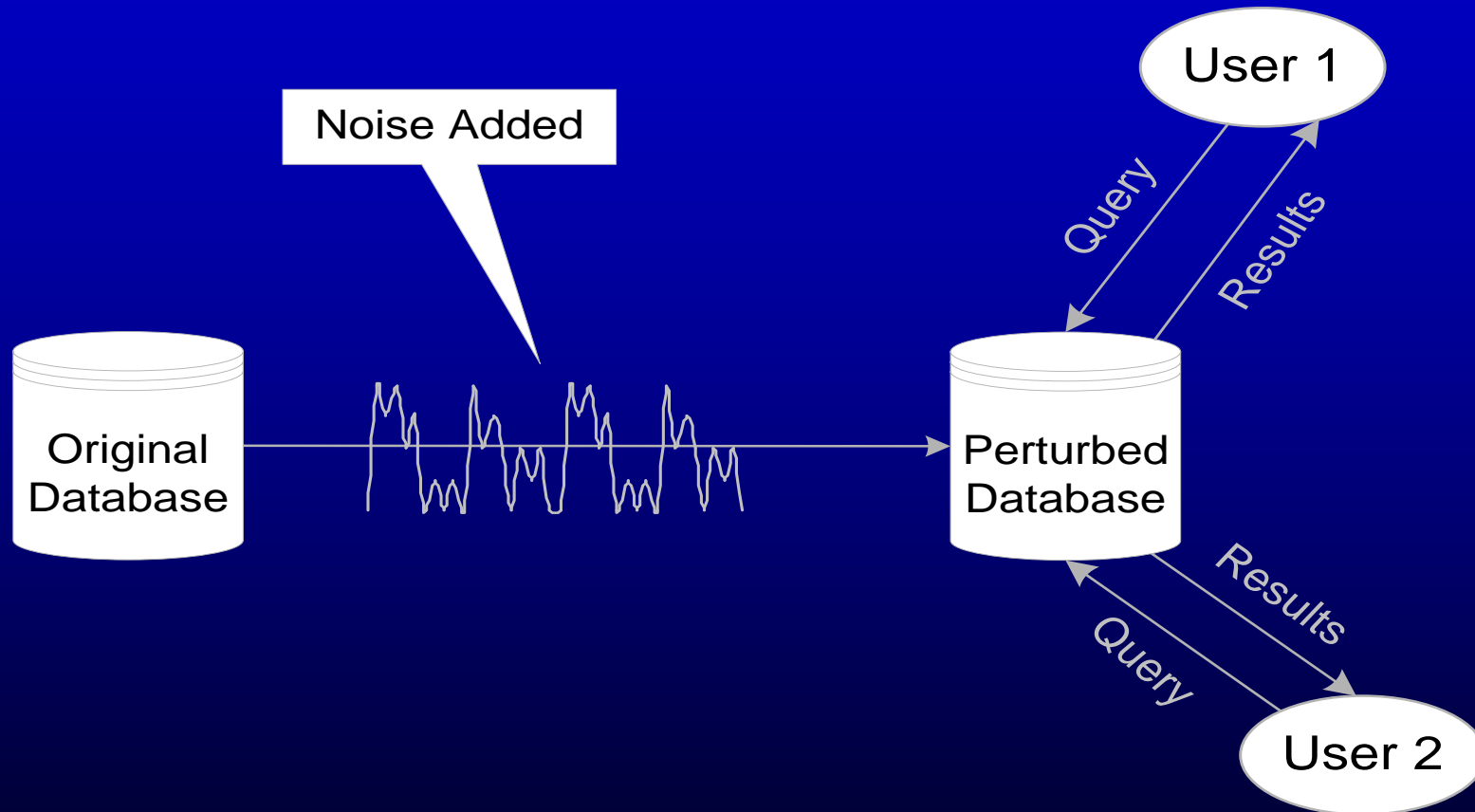
Các kỹ thuật chống suy diễn.



Kỹ thuật gây nhiễu dữ liệu

- *Gây nhiễu cố định (fixed perturbation)*
- *Gây nhiễu dựa vào truy vấn*

Data Perturbation



Kỹ thuật gây nhiễu dữ liệu...

- **Gây nhiễu cố định (fixed perturbation)**

- Cho N là kích cỡ của SDB và ta xét thuộc tính A_j .
- Mỗi giá trị thực x_{ij} (với $i=1, \dots, N$) của một thuộc tính A_j bị thay thế bằng một giá trị gây nhiễu x'_{ij}

$$x'_{ij} = x_{ij} + e_i \quad \text{với } i=1, \dots, N$$

- Vector $\mathbf{e} = (\mathbf{x}' - \mathbf{x}) = (e_1, \dots, e_N)$ là một vector gây nhiễu ngẫu nhiên
- $\mathbf{x} = (x_{1j}, \dots, x_{Nj})$, $\mathbf{x}' = (x'_{1j}, \dots, x'_{Nj})$ là các vector của giá trị thực và giá trị gây nhiễu của các bản ghi trong SDB, dành cho thuộc tính A_j

Kỹ thuật gây nhiễu dữ liệu...

- **Gây nhiễu cố định (fixed perturbation)**

- $\mathbf{e} = (\mathbf{e}_1, \dots, \mathbf{e}_N)$, mỗi thành phần \mathbf{e}_i là các biến ngẫu nhiên, độc lập tuyến tính.

$$E(\mathbf{e}_i) = 0, D(\mathbf{e}_i) = \sigma^2$$

- Các giá trị của mỗi thuộc tính A_j sẽ được cộng thêm một vector \mathbf{e} ngẫu nhiên.
 - Xác suất lỗi trong một câu truy vấn vượt quá giá trị giới hạn ε cho trước là:
 - $P(|q'(C) - q(C)| \geq \varepsilon |X(C)|) \leq \sigma^2 / (|X(C)| \varepsilon^2)$
 - Như vậy $|X(C)|$ càng lớn thì xác suất lỗi càng nhỏ

Kỹ thuật gây nhiễu dữ liệu...

- *Gây nhiễu cố định (fixed perturbation)*

- *Ưu điểm:*

- Chống được nhiễu tấn công, kể cả tấn công tính trung bình (lặp nhiều lần)

- *Nhược điểm:*

- Chỉ áp dụng cho thuộc tính số
 - Kết quả trả về không chính xác

Kỹ thuật gây nhiễu dữ liệu...

- *Gây nhiễu dựa vào truy vấn*
 - Không yêu cầu tạo một SDB nhiễu
 - Với mỗi truy vấn được tạo ra trong SDB, một *hàm gây nhiễu* sẽ được áp dụng với tất cả các thuộc tính của tập truy vấn đó.
 - Giả sử thống kê $q(C)$, với mọi giá trị x_{ij} thuộc $X(C)$: $x'_{ij} = f_c(x_{ij})$.
 - Giá trị $\varepsilon = x'_{ij} - x_{ij}$ là ngẫu nhiên.

Kỹ thuật gây nhiễu dữ liệu...

- **Gây nhiễu dựa vào truy vấn**

- **Thống kê Sum:**

- Xét thống kê **$S = q(C) = \text{Sum}(C, A_j)$** , **$n$** là số lượng các bản ghi tập truy vấn $X(C)$.

- **$S' = \sum_{i=1}^n x'_{ij}$** với **$x'_{ij} = f(x_{ij}) = x_{ij} + z_1 (x_{ij} - \overline{x_{C_j}}) + z_2$**

- **z_1** và **z_2** là các biến ngẫu nhiên độc lập được sinh ra cho mỗi bản ghi

Kỹ thuật gây nhiễu dữ liệu...

- *Gây nhiễu dựa vào truy vấn*

- *Thống kê Count:*

- Giả sử thống kê $\text{Count}(C) = m$

- $m' = \sum_{j=3}^n z_3$

Với $E(z_3) = 1$ và $\text{Var}(z_3) = a^2_1 / m$,

- và z_3 được sinh ngẫu nhiên và độc lập với các bản ghi x_i trong $X(C)$.

- $E(m') = m$ và $\text{Var}(m') = a^2_1$

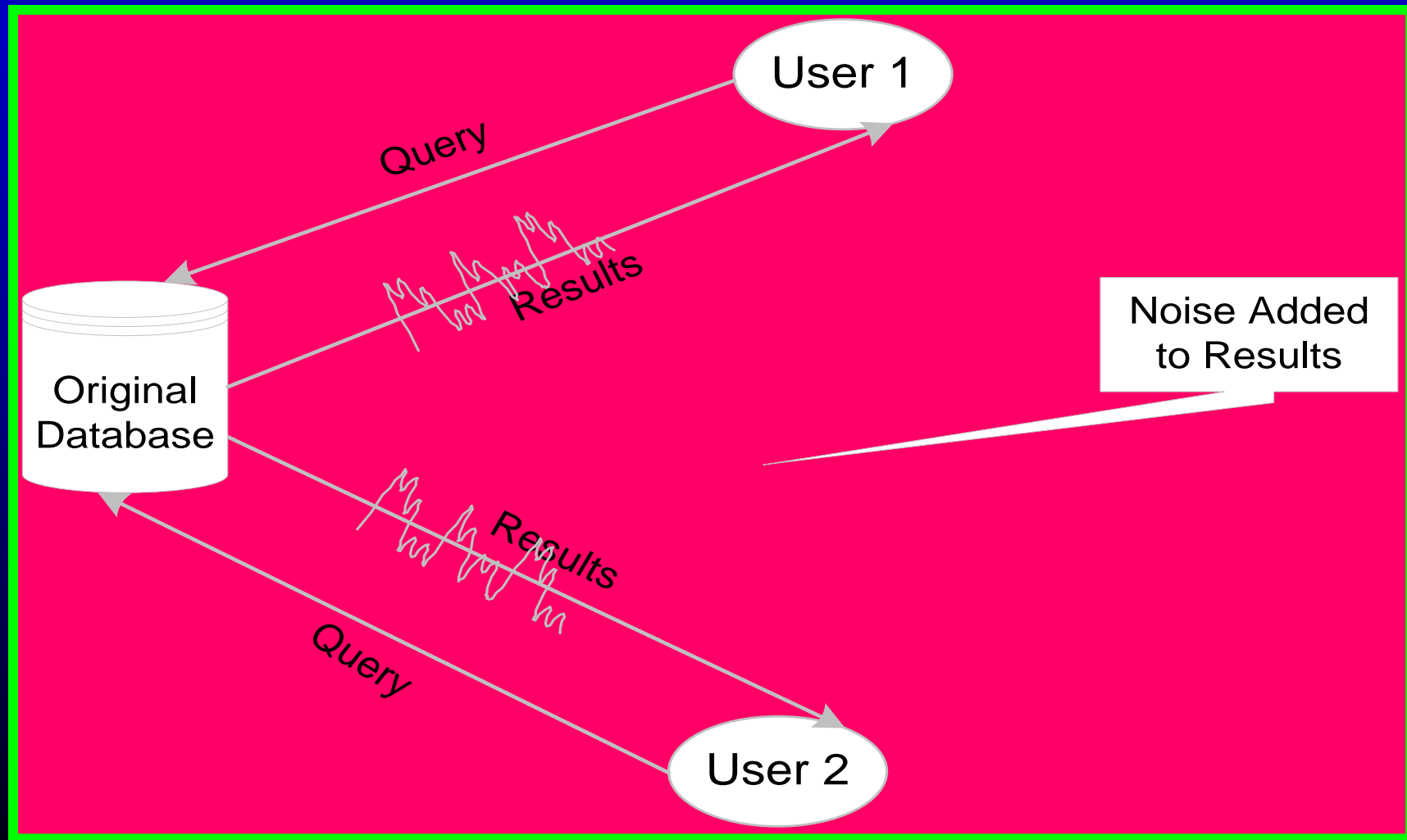
Kỹ thuật gây nhiễu dữ liệu...

- *Gây nhiễu dựa vào truy vấn*
- *Ưu điểm:*
 - Gây nhiễu dữ liệu nên chống được nhiều tấn công
- *Nhược điểm:*
 - Với mỗi thống kê, lại phải áp dụng một hàm gây nhiễu f , với giá trị nhiễu \Rightarrow tốn công, giảm hiệu năng hệ thống.
 - Kết quả đưa ra không chính xác.

Các kỹ thuật chống suy diễn.



Kỹ thuật gây nhiễu đầu ra



Kỹ thuật gây nhiễu đầu ra...

- Các *kỹ thuật gây nhiễu đầu ra* thực hiện sửa đổi trên các kết quả được tính toán chính xác của một câu truy vấn thống kê, trước khi chuyển nó cho người sử dụng.
- *Kỹ thuật Làm tròn (rounding)*

Kỹ thuật gây nhiễu đầu ra...

- *Kỹ thuật Làm tròn (rounding)*

- Kết quả mọi câu truy vấn sẽ được làm tròn:

$$Q' = r(Q)$$

- *Làm tròn có hệ thống (systematic rounding)*

- *Làm tròn ngẫu nhiên (random rounding)*

Kỹ thuật gây nhiễu đầu ra...

- **Làm tròn có hệ thống (systematic rounding)**
 - **Q'** là một kết quả sửa đổi, nó được tính toán cho thống kê yêu cầu $q(C)$.
 - **$b' = \lfloor (b+1)/2 \rfloor$** (ký hiệu $\lfloor \rfloor$ chỉ làm tròn xuống số nguyên gần nhất), giá trị b do Admin chọn.
 - **$d = Q \bmod b$.**

$$- \mathbf{r(Q) = \begin{cases} Q & \text{nếu } d = 0 \\ Q - d & \text{nếu } d < b' \\ Q + b - d & \text{nếu } d \geq b' \end{cases}}$$

Kỹ thuật gây nhiễu đầu ra...

- **Làm tròn ngẫu nhiên (random rounding)**

- **Q'** là một kết quả sửa đổi, nó được tính toán cho thống kê yêu cầu $q(C)$.
- **$b' = \lfloor (b+1)/2 \rfloor$** (ký hiệu $\lfloor \rfloor$ chỉ làm tròn xuống số nguyên gần nhất)
- **$d = Q \bmod b$.**

$$- \mathbf{r(Q)} = \begin{cases} Q & \text{nếu } d = 0 \\ Q - d & \text{với xác suất } 1 - p \\ Q + b - d & \text{với xác suất } p \end{cases}$$

Xác suất **$p = d/b$**

Kỹ thuật gây nhiễu đầu ra...

- *Kỹ thuật Làm tròn (rounding)*
- *Ưu điểm:* Bảo vệ được những tấn công đơn giản.
- *Nhược điểm:*
 - Không chống được những tấn công trung bình, tấn công trình theo dõi
 - Kết quả đưa ra cũng không chính xác.

Các kỹ thuật chống suy diễn.



Kỹ thuật mẫu ngẫu nhiên

- Cục điều tra dân số Mỹ sử dụng **kỹ thuật mẫu ngẫu nhiên** để ngăn chặn suy diễn trong các cơ sở dữ liệu thống kê.
- **Ý tưởng:** của kỹ thuật này là sử dụng các mẫu bản ghi từ các tập truy vấn tương ứng với các truy vấn thống kê, thay vì lấy mẫu trong toàn bộ SDB.

Kỹ thuật mẫu ngẫu nhiên

Giả thiết

- Công thức đặc trưng C
- Tập truy vấn $X(C)$
- Thống kê trên C : $q(C)$

Phương pháp

Thay vì tính $q(C)$ trên tập $X(C)$, ta tính trên một **mẫu ngẫu nhiên** gồm m bản thi trong $X(C)$

- $m < |X(C)|$

Kỹ thuật mẫu ngẫu nhiên...

- Cơ chế cơ bản của kỹ thuật này là thay thế tập truy vấn (có liên quan đến một câu truy vấn thống kê) bằng một tập truy vấn được lấy mẫu (**sampled query set**) gồm một tập con các bản ghi được chọn lựa chính xác trong tập truy vấn gốc.
- Sau đó, tiến hành tính toán thống kê yêu cầu trên tập truy vấn mẫu này. Sử dụng một hàm chọn $f(C, i)$ để chọn lựa các bản ghi từ tập truy vấn gốc tương ứng với thống kê $q(C)$ mà người dùng yêu cầu.

So sánh các kỹ thuật chống suy diễn

- **Các tiêu chuẩn so sánh:**

- **Security:** đánh giá mức độ bảo vệ của kỹ thuật (chống được những tấn công nào), chống được suy diễn, có lộ chính xác, lộ từng phần không.
- **Mức đầy đủ của thông tin:** kết quả trả về có chính xác không, có nhất quán không và có bị mất mát thông tin hay không.
- **Cost:** chi phí thực hiện, chi phí xử lý trên một câu truy vấn (thời gian CPU), chi phí đào tạo người dùng.

So sánh các kỹ thuật chống suy diễn

Method	Security	Richness of Information	Costs
Query-set Restriction	Low	Low ¹	Low
Microaggregation	Moderate	Moderate	Moderate
Data Perturbation	High	High-Moderate	Low
Output Perturbation	Moderate	Moderate-low	Low
Auditing	Moderate-Low	Moderate	High
Sampling	Moderate	Moderate-Low	Moderate

Xin chân thành cảm ơn!