CHƯƠNG 5. AN TOÀN CƠ SỞ DỮ LIỆU THỐNG KÊ

TS. Trần Thị Lượng
* Khoa An toàn thông tin *

NỘI DUNG

1. Kỹ thuật khái niệm
2. Kỹ thuật hạn chế
3. Kỹ thuật gây nhiễu
4. Kỹ thuật mẫu ngẫu nhiên

MỤC TIÊU BÀI HỌC

❑ Kiến thức

- Hiểu được cơ chế hoạt động của các kỹ thuật kiểm soát suy diễn.

❑ Kỹ năng

- Thực hiện tấn công Trình theo dõi và Hệ tuyến tính vào CSDL thống kê, và sử dụng các kỹ thuật kiểm soát suy diễn để chống các tấn công suy diễn trên.

TÀI LIỆU THAM KHẢO

1. Giáo trình "*An toàn cơ sở dữ liệu*"// Chương 4 "An toàn cơ sở dữ liệu thống kê"
2. Nabil R.Adam, John C. Wortmann, *Security-Control Methods for Statistical Databases: A comparative Study*, ACM Computing Surveys, Vol. 21, No.4, December 1989.
3. Shiuh-Pyng Shieh, Chern-Tang Lin, *Information Protection in Dynamic Statistical,* National Chiao Tung University Databases.

TÀI LIỆU THAM KHẢO

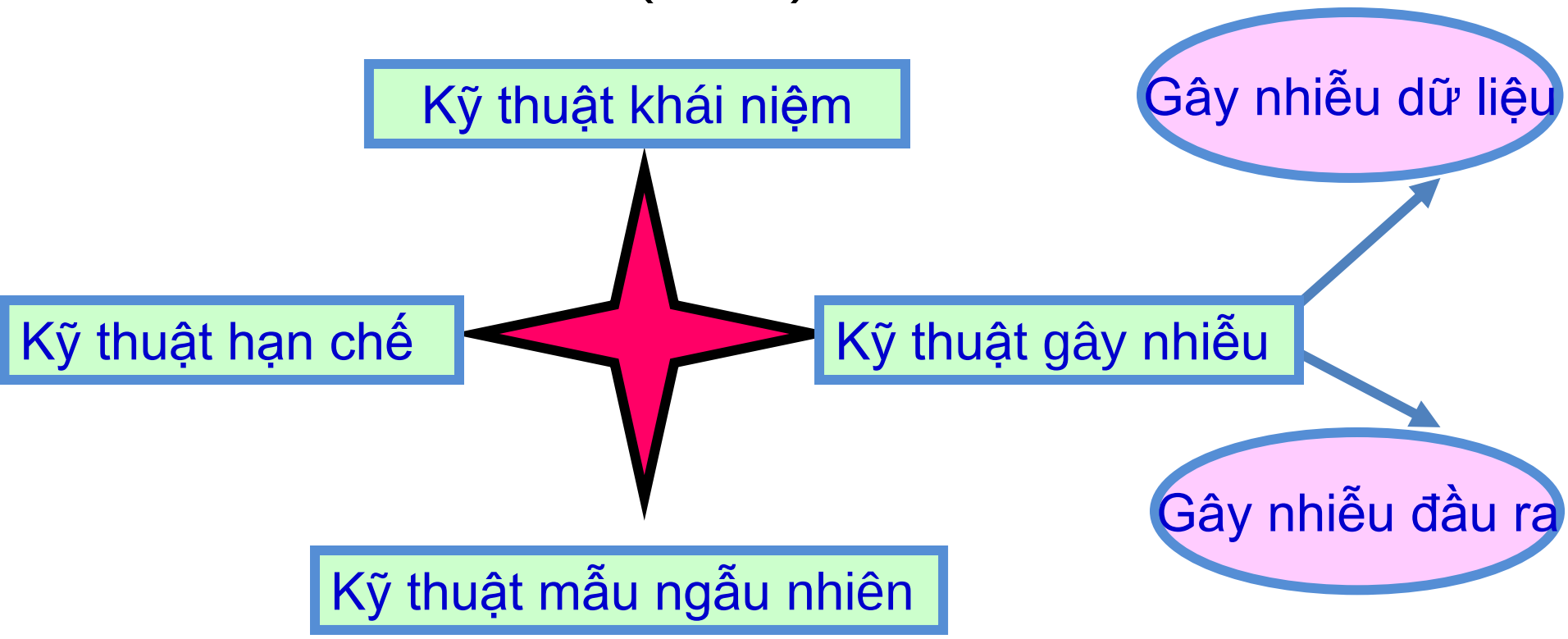
- Website:
  - *Tổng cục thống kê – Việt Nam:* http://www.gso.gov.vn/default.aspx?tabid=228&ItemID=1915
  - *SDB Liên hợp quốc (UNO)* http://unstats.un.org/unsd/databases.htm
  - *SDB kinh tế khối Châu Âu (UNECE)* http://w3.unece.org/pxweb/Dialog/
  - *SDB WTO* http://stat.wto.org/Home/WSDBHome.aspx?Language=
  - v.v

NỘI DUNG

1. Kỹ thuật khái niệm
2. Kỹ thuật hạn chế
3. Kỹ thuật gây nhiễu
4. Kỹ thuật mẫu ngẫu nhiên

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



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 năm 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, v.v.

Ví dụ: MÔ HÌNH LƯỚI CHO SDB VỀ CÔNG NHÂN

Dạng vĩ mô

COUNT

Bảng 3-chiều (N=3)

Năm sinh	Giới tính	Mã phòng		
		Phong1	Phong2	Phong3
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…

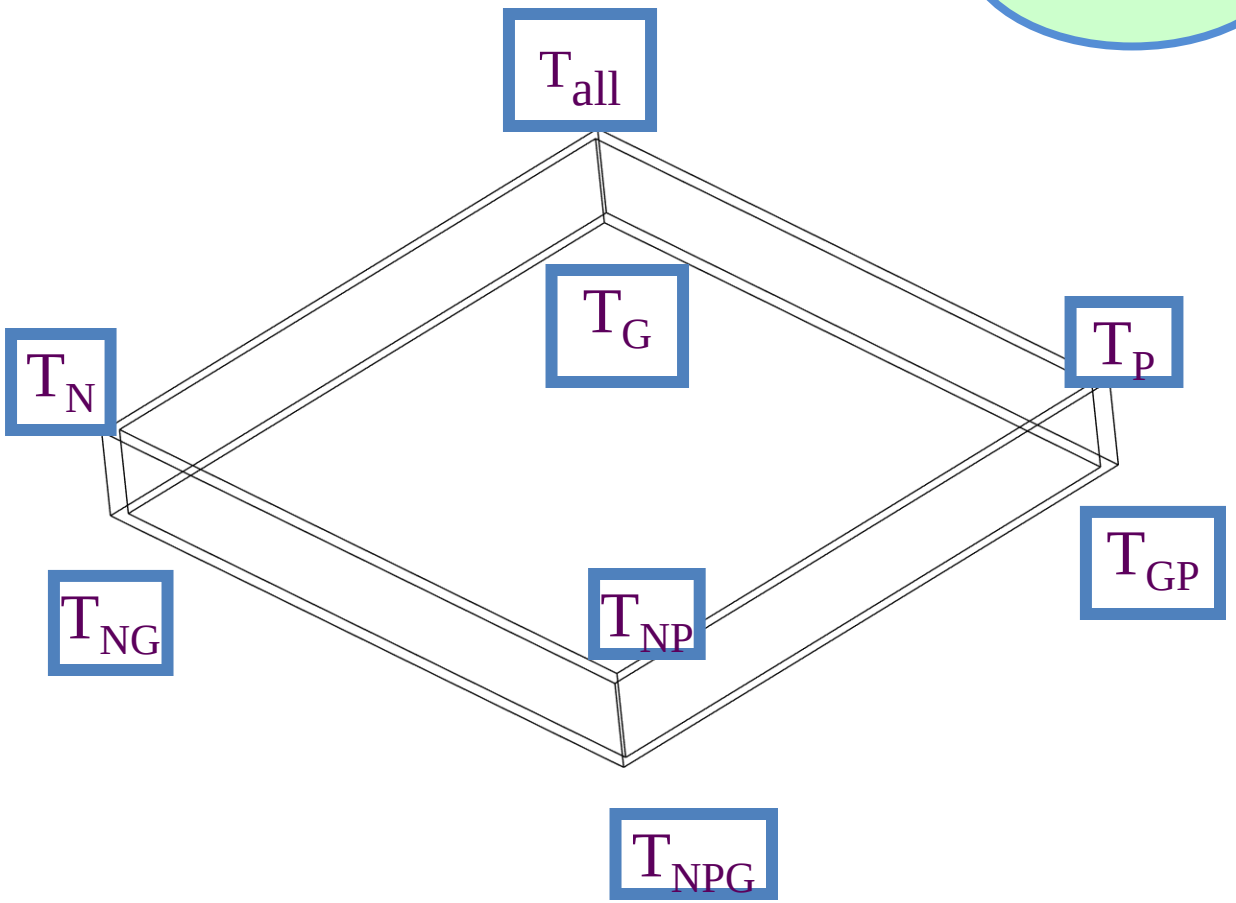
- *Cấu trúc lưới:*

COUNT

N: Năm sinh

G: Giới tính

P: Mã phòng



KỸ THUẬT KHÁI NIỆM…

- *Các bảng 2-chiều*

NG Table

Năm	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:*

129

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.

KỸ THUẬT KHÁI NIỆM

- Cấu trúc lưới:
  - Ưu điểm:

+ 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.

KỸ THUẬT KHÁI NIỆM

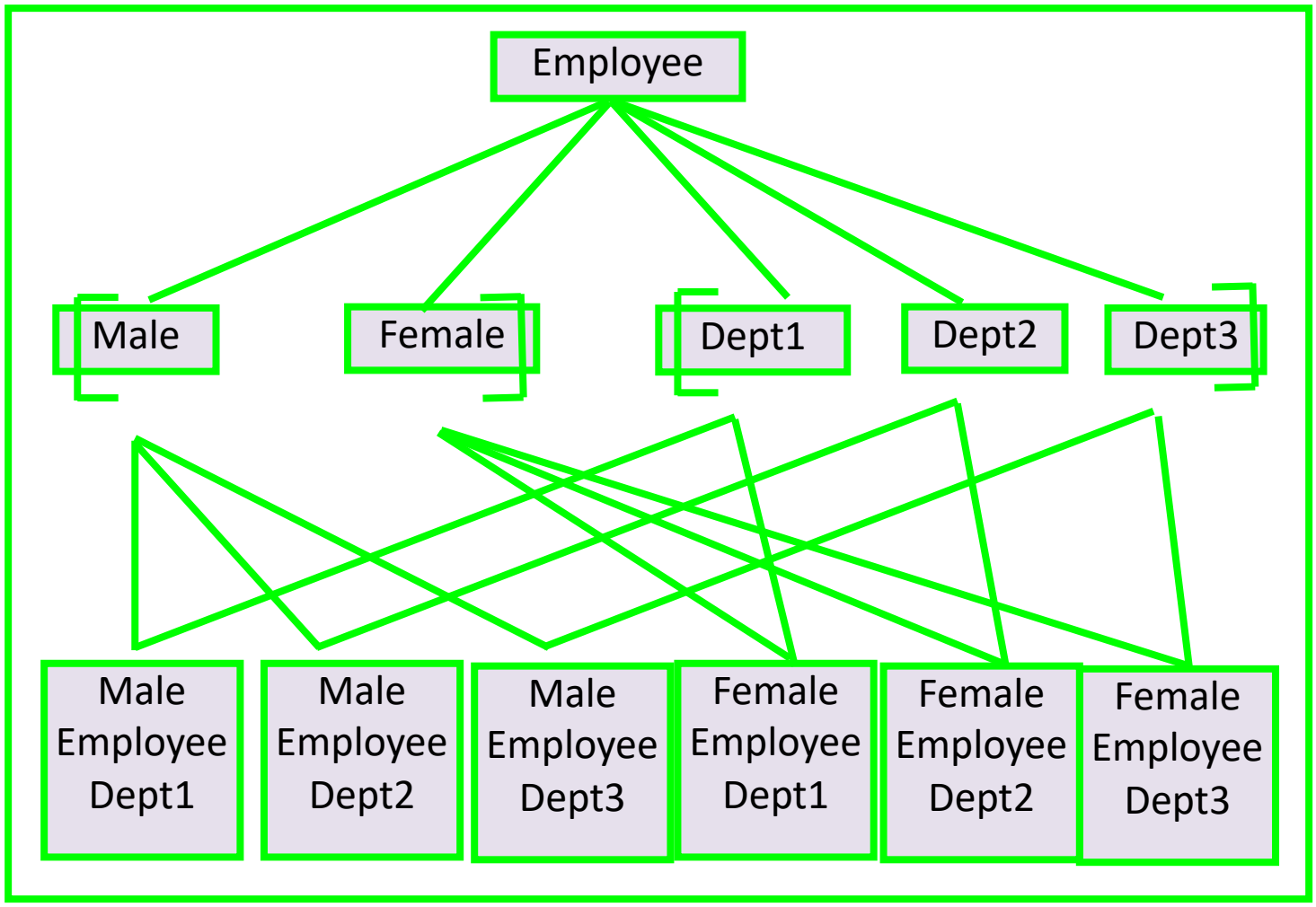
- *Phân hoạch khái niệm:*
  - Dựa vào việc định nghĩa tập các cá thể của SDB tại mức khái niệm, được gọi là các *lực lượng* (*populations*).
  - Dựa vào các điều kiện cần kiểm tra nhằm tránh suy diễn

KỸ THUẬT KHÁI NIỆM

- *Phân hoạch khái niệm:*
  - Hình sau minh hoạ 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*, tuỳ 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:*



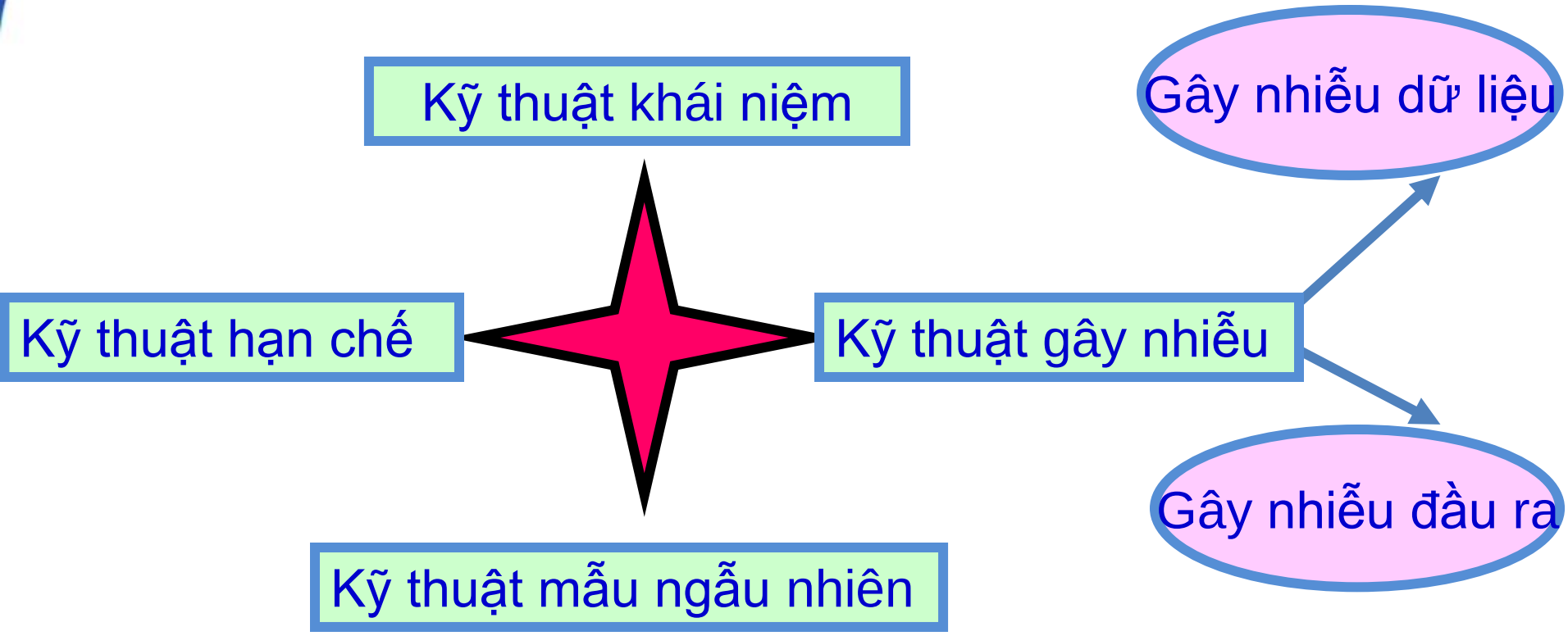
KỸ THUẬT KHÁI NIỆM

- *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)

NỘI DUNG

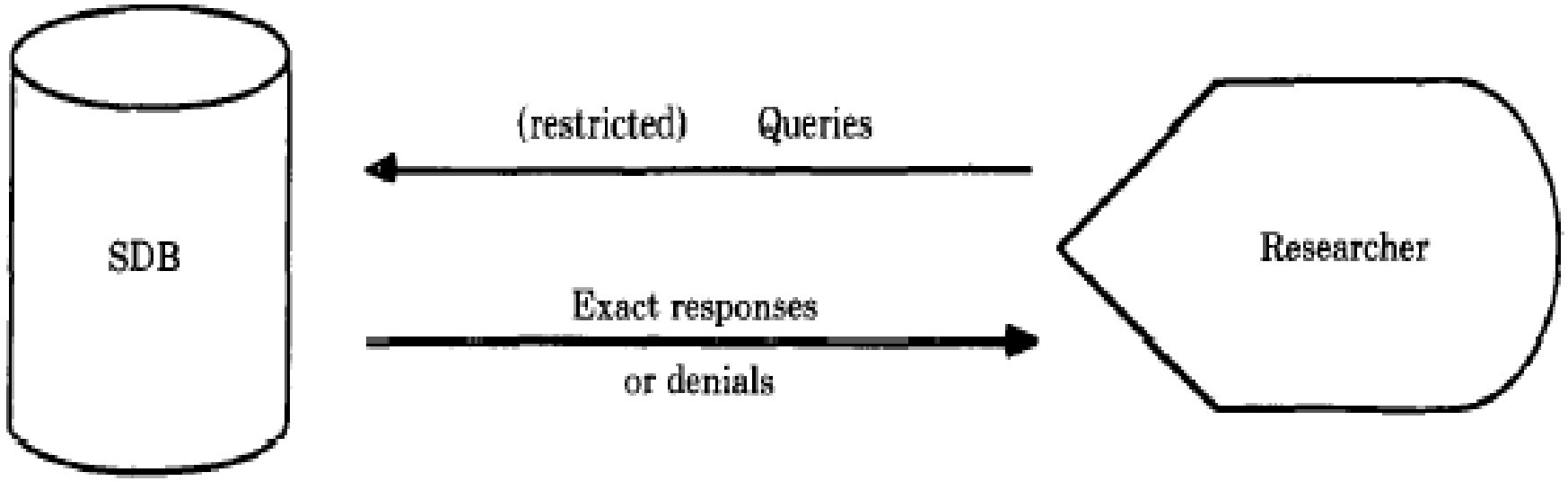
1. Kỹ thuật khái niệm
2. Kỹ thuật hạn chế
3. Kỹ thuật gây nhiễu
4. Kỹ thuật mẫu ngẫu nhiên

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

- 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

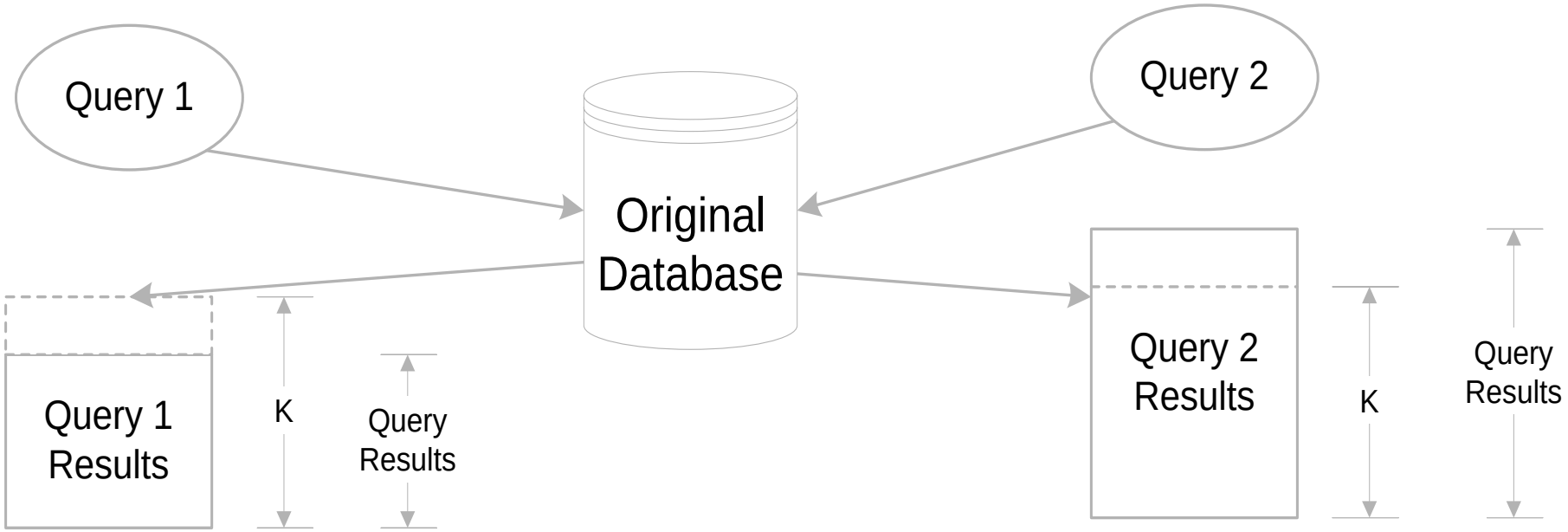
KIỂM SOÁT KÍCH CỠ TẬP TRUY VẤN

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

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

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

KIỂM SOÁT KÍCH CỠ TẬP TRUY VẤN



| X(C) | < K

NO

| X(C) | > K

YES

KIỂM SOÁT KÍCH CỠ TẬP TRUY VẤN

- 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	Tài vụ	24	F	2900

KIỂM SOÁT KÍCH CỠ TẬP TRUY VẤN

N = 5, chọn K =2

$K \le |X(C)| \le N-K$

$0 \le K \le N/2$

- Công thức đặc trưng C1

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

- 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)

KIỂM SOÁT KÍCH CỠ TẬP TRUY VẤN

- 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	Tài vụ	24	F	2900

KIỂM SOÁT KÍCH CỠ TẬP TRUY VẤN

N = 5, Chọn K =2

$K \le |X(C)| \le N-K$

$0 \le K \le N/2$

- Công thức đặc trưng C2

C2 = (ChucVu = "Giám sát viên")

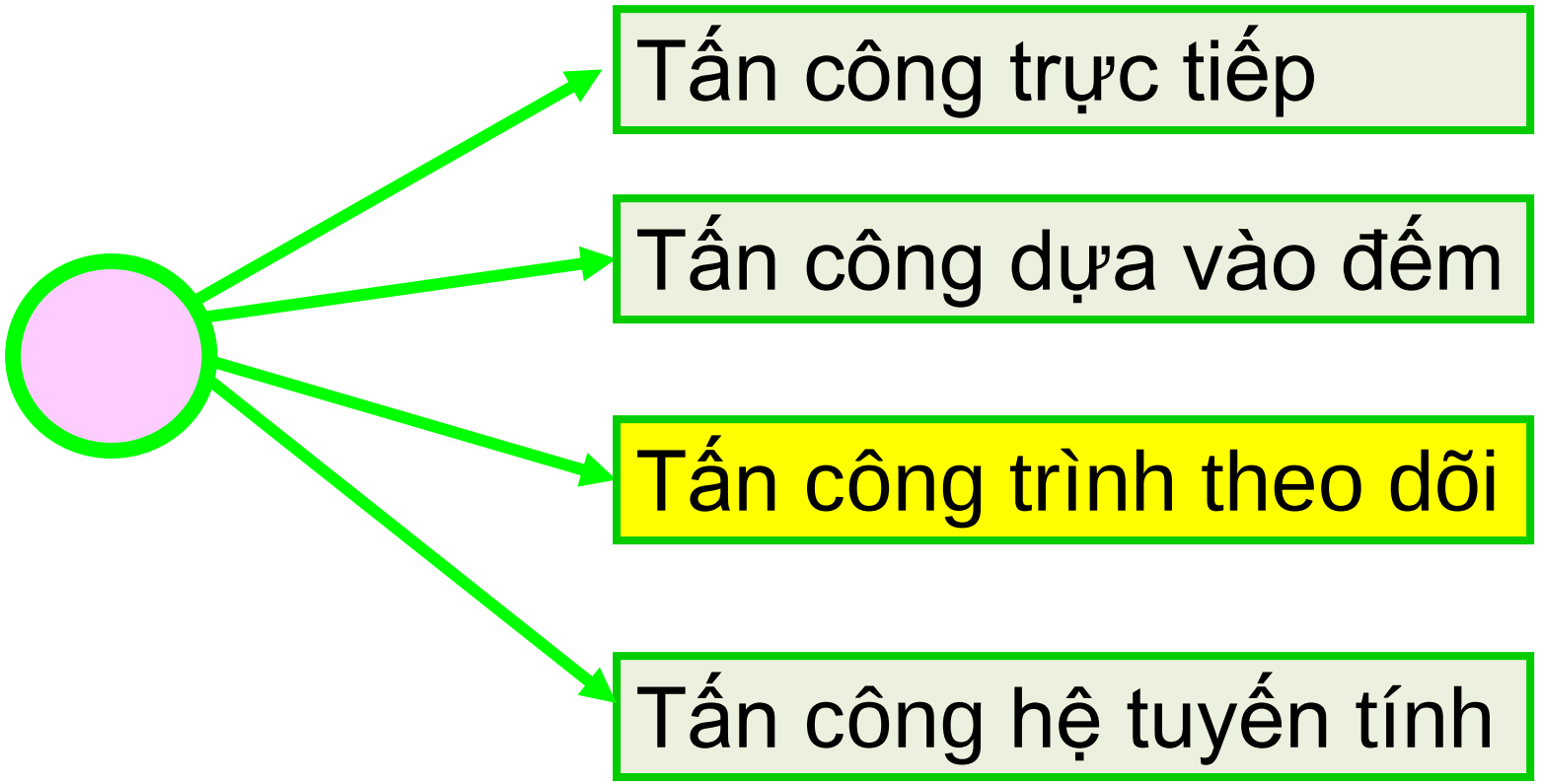
- Tập truy vấn X(C2)

ID	Ten	ChucVu	PhongLV	GioiTinh	Lương
04	Minh	Giám sát viên	Maketing	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



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 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 DỰA VÀO 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:

$$K \leq |X(T)| \leq N\text{-}K$$

TẤN CÔNG DỰA VÀO TRÌNH THEO DÕI

Kiểu 1

- *Giả thiết:*
  - User cần tính Count(C), Sum(C, Luong)
  - Công thức $C = (A \wedge B)$, và **Count (C) = 1**.
    => *Count(C), Sum(C, Luong) bị cấm!*
- *Tấn công:*
  - Bước 1: Tính T = $A \wedge \bar{B}$ thỏa mãn $k \leq |X(T)| \leq N-k$.
  - Bước 2: Tính gián tiếp Q(C) bằng:
    $Q(C) = Q(A \wedge B) = Q(A) - Q(A \wedge \bar{B})$
    => $Q(C) = Q(A) - Q(T)$

TẤN CÔNG DỰA VÀO TRÌNH THEO DÕI

- 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	Tài vụ	24	F	2900

TẤN CÔNG DỰA VÀO TRÌNH THEO DÕI

Giả thiết:

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

- User cần tính *Count(C), Sum(C, Lương)*
- Count (C) = 1. ⇨ ***Các câu truy vấn này bị cấm!***

N=5

K=2

- **Tấn công:**

+ Đặt C=(A∧ B)

A = (Phong='Kế hoạch')

B = (Tuoi =24) ∧(GioiTinh = F)

+ Tính Count(T) = Count(A∧$\bar{B}$) =2 thỏa mãn

$2 \leq$ Coun(T)=2 ≤ 3.

+ Tính gián tiếp Count (C):

Count(C)= Count (A∧B)

= Count(A)-Count(A∧$\bar{B}$)

Count(C) = Count(A) - Count(T) = 3 -2 =1

TẤN CÔNG DỰA VÀO TRÌNH THEO DÕI

- Tính Sum(C, Luong):

+ Đặt C=(A∧ B)

A = (Phong=‘Kế hoạch’)

B = (Tuoi =24) ∧(GioiTinh = F)

+ Tính gián tiếp Sum(C,Luong):

Sum(C,Luong)= Sum (A∧B, Luong)

= Sum(A,Luong)-Sum(A∧ $\bar{B}$,Luong)

Sum(C,Luong) = (6200+4000+ 2900) – (6200+4000) = 2900

➡ Đây chính là lương của nhân viên Quỳnh

BÀI TẬP 1: TẤN CÔNG DỰA VÀO TRÌNH THEO DÕI KIỂU 1

NhanVien

ID	Ten	ChucVu	Phong	Tuoi	GioiTinh	Luong
01	Nam	Nhân viên	Maketing	29	F	3500
02	Lan	Trưởng phong	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 = “Nam”)$\wedge$ (ChucVu = “Phó phòng”)

TẤN CÔNG DỰA VÀO TRÌNH THEO DÕI

Kiểu 2

- Giả thiết:
  - *Cần tính Count(C), Sum(C,A_j), mà Count(C)<K*
  - *Các thống kê này bị cấm!*
- *Tấn công:*
  - Bước 1: Chọn T thỏa mãn **k≤ |X(T)|, | X($\bar{T}$) | ≤ N-k.**
  - Bước 2: Tính **Q(D)= Q(All) = Q(T) + Q($\bar{T}$)** (Q(All) bị cấm)
  - Bước 3: Tính gián tiếp Q(C) bằng:

$$Q(C) = Q(CvT) + Q(Cv\bar{T}) - Q(D)$$

TẤN CÔNG DỰA VÀO TRÌNH THEO DÕI

- Ví dụ SDB về các vụ tai nạn môtô

HoTen	Tuoi	Đ/C	MauXe	LoaiXe	ThoiGian	CoLoi	SayRuou
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

TẤN CÔNG DỰA VÀO TRÌNH THEO DÕI

Kiểu 2

- ***Giả thiết:*** C = (Ten='Minh') ∧(MauXe='Trắng')
  - Count(C)=1, SUM(CoLoi, C)=1

  2 Câu truy vấn này bị cấm!
- ***Tấn công:***
  - Chọn **T = (Tuoi<25)** => Count(T)=2, Count($\bar{T}$)=3
  - Count(All)= Count(T) + Count($\bar{T}$) =5
  - Tính:

  + Count(C) = Count(C v T) + Count(C v $\bar{T}$)–Count(All)
  = 3 + 3 – 5 = 1

  + SUM(CoLoi, C)= Sum(CoLoi, C v Tuoi<25) + Sum(CoLoi, C v Tuoi>=25) – Sum(CoLoi, All)
  = 2 + 3 – 4 = 1.

Anh Minh có lỗi trong vụ tai nạn đó!

BÀI TẬP 2: TẤN CÔNG DỰA VÀO TRÌNH THEO DÕI KIỂU 2

NhanVien

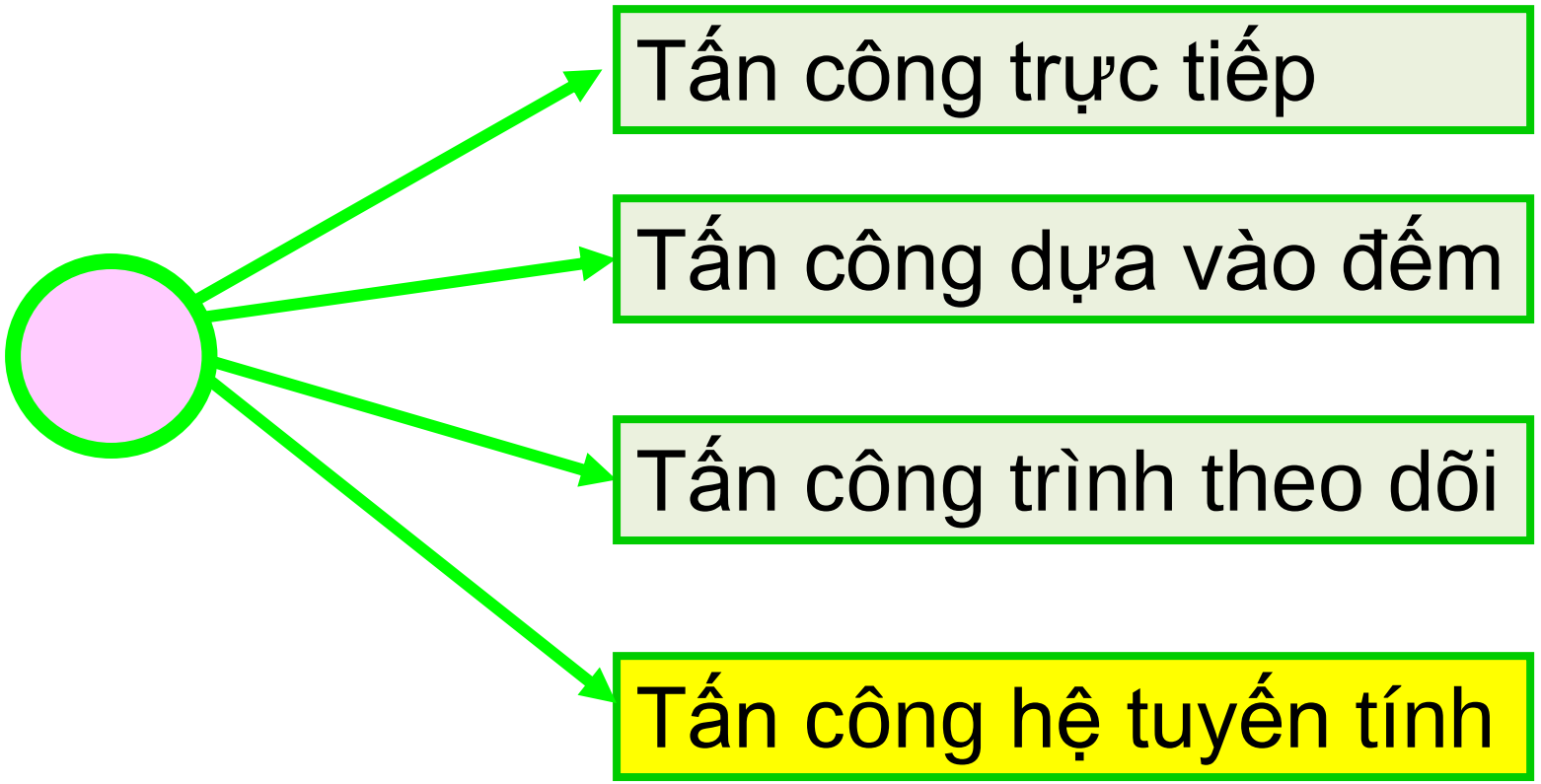
ID	Ten	ChucVu	Phong	Tuoi	GioiTinh	Luong
01	Nam	Nhân viên	Maketing	29	F	3500
02	Lan	Trưởng phong	Kế hoạch	33	M	6200
03	Huệ	Nhân viên	Kế hoạch	27	F	4000
04	Minh	Giám sát viên	Maketing	24	F	3600
05	Nam	Nhân viên	Kế hoạch	24	M	2900
06	Yến	Nhân viên	Tài vụ	40	F	4600
07	Nam	Phó phòng	Tài vụ	38	M	5000

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

KIỂM SOÁT KÍCH CỠ TẬP TRUY VẤN

- Ưu điểm:
  - Đưa ra kết quả chính xác
  - Chỉ chống được các tấn công đơn giản
- 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.*

KIỂM SOÁT KÍCH CỠ TẬP TRUY VẤN



TẤN CÔNG HỆ TUYẾN TÍNH

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

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

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

$$\vdots$$

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

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

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

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

TẤN CÔNG HỆ TUYẾN TÍNH

- **Ví dụ 1:**

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	Tài vụ	24	F	2900

TẤN CÔNG HỆ TUYẾN TÍNH

- *Giả thiết:*
  - *C = (Phong='Kế hoạch')∧(GioiTinh=F)*
  - Cần tính q= Count(C) =1 ⇨ Bị chặn!
- Thực hiện:
  - Tính q_1 = Count(Phong='Kế hoạch')
  - Tính q_2 = Count(Phong='Kế hoạch', GioiTinh =M)

⇨

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

$\Rightarrow q_3$= Count(Phong='Kế hoạch', GioiTinh =F)

$= q_1 - q_2 = 3 - 2 = 1.$

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

TẤN CÔNG HỆ TUYẾN TÍNH

- ***C = (Phong='Kế hoạch') ∧(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.

- **Ví dụ 2:**
- Giả sử cần tính q_3= *Sum*(*Sex* = *M* $\wedge$ *Dept-Code* = *Dept3*$\wedge$ *Birth-Year* = 1968, *Salary*), count = 1.
- Ta có:

$$\begin{cases} q1 = \text{Sum}\ (\text{Sex} = F \wedge \text{Dept - Code} = \text{Dept3} \wedge \text{Birth - Year} = 1968, \text{Salary}) \\ q2 = \text{Sum}\ ((\text{Sex} = F \vee \text{Sex} = M) \wedge \text{Dept - Code} = \text{Dept3} \wedge \text{Birth - Year} = 1968, \text{Salary}) \end{cases}$$

  - Tương ứng ta có hệ sau:
  - Count1 = 7
  - Count2 = 8

$$\begin{cases} x1 + x3 + x4 + x6 + x7 + x8 + x9 = 33 \\ x1 + x3 + x4 + x5 + x6 + x7 + x8 + x9 = 37 \end{cases}$$

  => Tính được: $x5 = q_2 - q_1 = 4$.

  => Và người dùng biết count (q_3) = 1 => có thể tìm được lương của người này

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 GIẤU Ô(Cell suppression)

- Tài liệu tham khảo:
  - [1]. Matteo Fischetti • Juan José Salazar, *Solving the Cell Suppression Problem on Tabular Data with Linear Constraints*, Management Science © 2001 INFORMS Vol. 47, No. 7, July 2001 pp. 1008–1027.
  - [2]. James P. Kelly, Bruce L. Golden, Arjang A. Assad, *Cell suppression: Disclosure protection for sensitive tabular data,* Networks journal, Volume 22, Issue 4 July 1992, Pages 397–417

KỸ THUẬT GIẤU Ô

- 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

- Tiêu chuẩn giấu ô:
  - *Thống kê Count:* kích cỡ tập truy vấn nhỏ hơn hoặc bằng 1, nghĩa là Count(C) = 0, 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" => ô này bị giấu*
    - Các tham số **n** và **k** được giữ bí mật và do DBA xác định (n<N)

KỸ THUẬT GIẤU Ô

ID	Ten	ChucVu	Phong	Tuoi	GioiTinh	Lương
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

KỸ THUẬT GIẤU Ô

- **Ví dụ 1:** 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 hoach	Maketing	Tài vụ	
<27 27-30 >30	4500(2) 9000(2) 7000 (2)	3600(1) 3500(1) 11200(2)	3700 (2) 0 (0) 8100(3)	11800 12500 27200
Tổng lương	20500	18300	12700	51500

- **Ví dụ 2:** Giả sử ***n* = 2** và ***k* = 90%**

SUM

Địa chỉ	Mã phòng			Tổng lương
	Phong1	Phong 2	Phong 3	
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

- Giả sử kết quả giấu ô như sau:

Ví dụ 2

Địa chỉ	Mã phòng			Tổng lương
	Phong1	Phong 2	Phong 3	
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

- Giả sử kết quả giấu ô như sau:

Ví dụ 2

Địa chỉ	Mã phòng			Tổng lương
	Phong1	Phong 2	Phong 3	
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

Cần giấu ô bổ sung như thế nào?

KỸ THUẬT GIẤU Ô

- Cần giấu ô bổ sung như sau:

Địa chỉ	Mã phòng			Tổng lương
	Phong1	Phong 2	Phong 3	
Hà Nội	135	80	50	265
Hải Phòng	----	360	----	580
Nam Định	225	90	900	1215
Nghệ An	----	210	----	585
Tổng lương	780	740	1125	2645

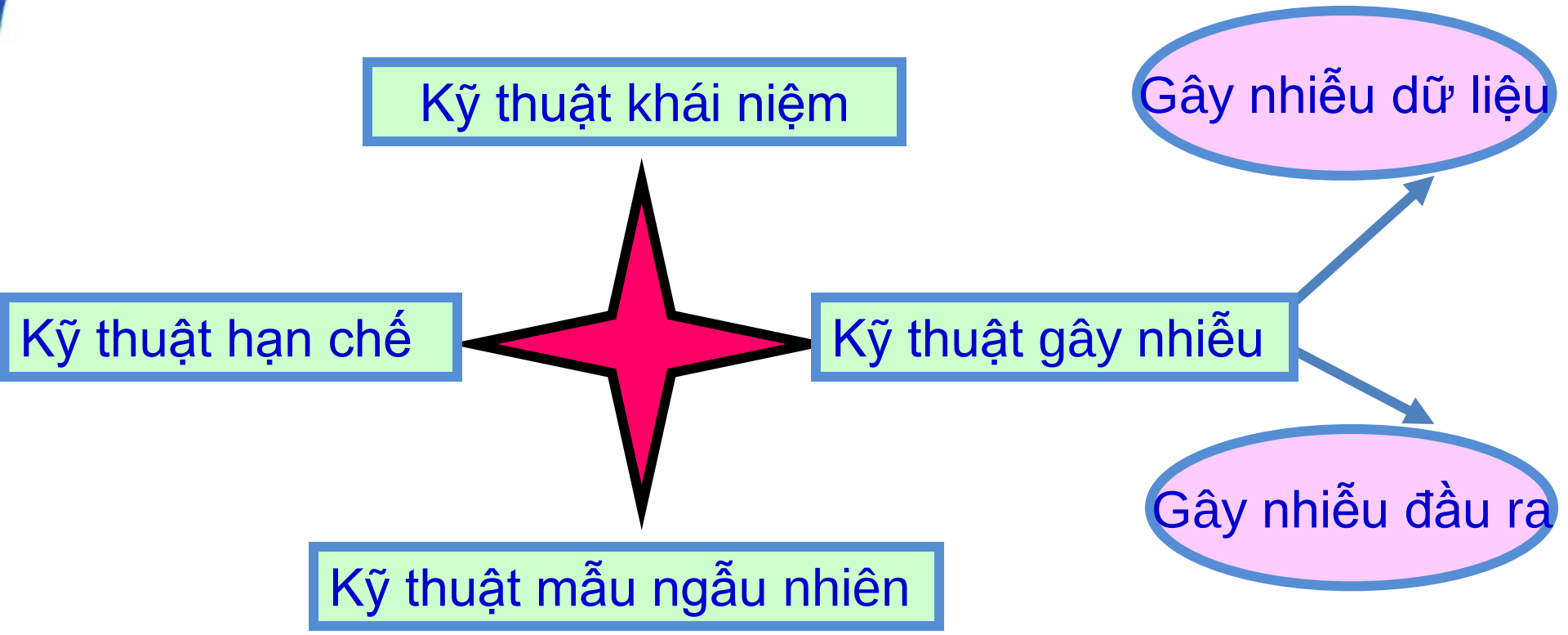
KỸ THUẬT GIẤU Ô

- 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.

NỘI DUNG

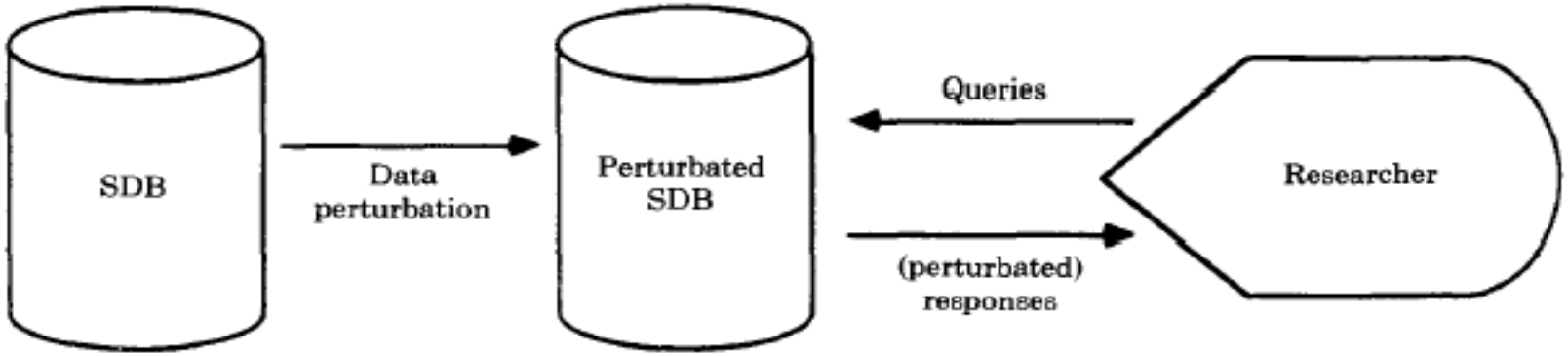
1. Kỹ thuật khái niệm
2. Kỹ thuật hạn chế
3. Kỹ thuật gây nhiễu
4. Kỹ thuật mẫu ngẫu nhiên

CÁC KỸ THUẬT CHỐNG SUY DIỄN.

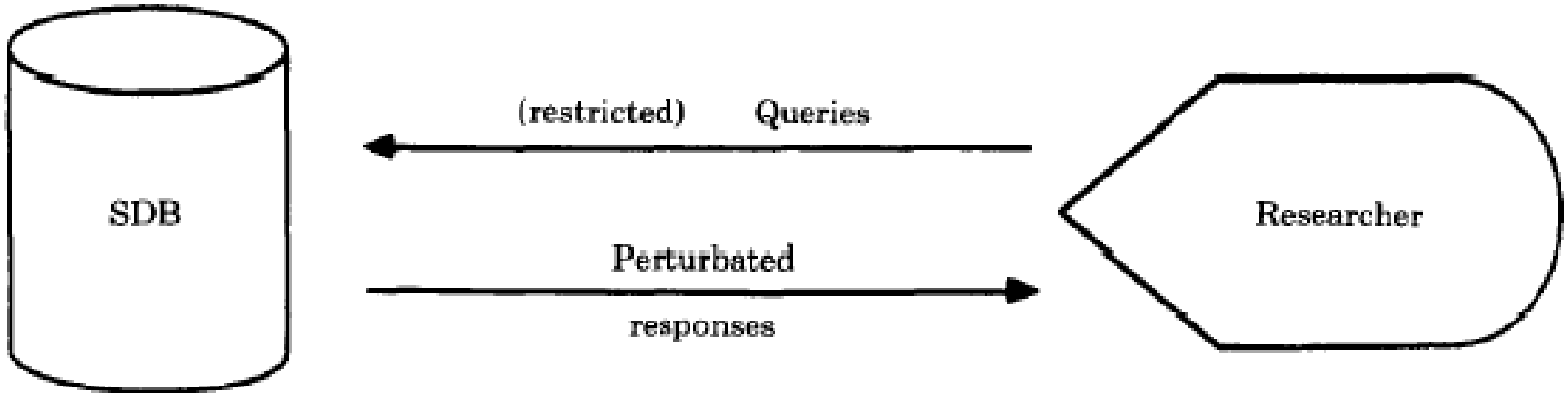


CÁC KỸ THUẬT DỰA VÀO GÂY NHIỄU

- Kỹ thuật gây nhiễu dữ liệu

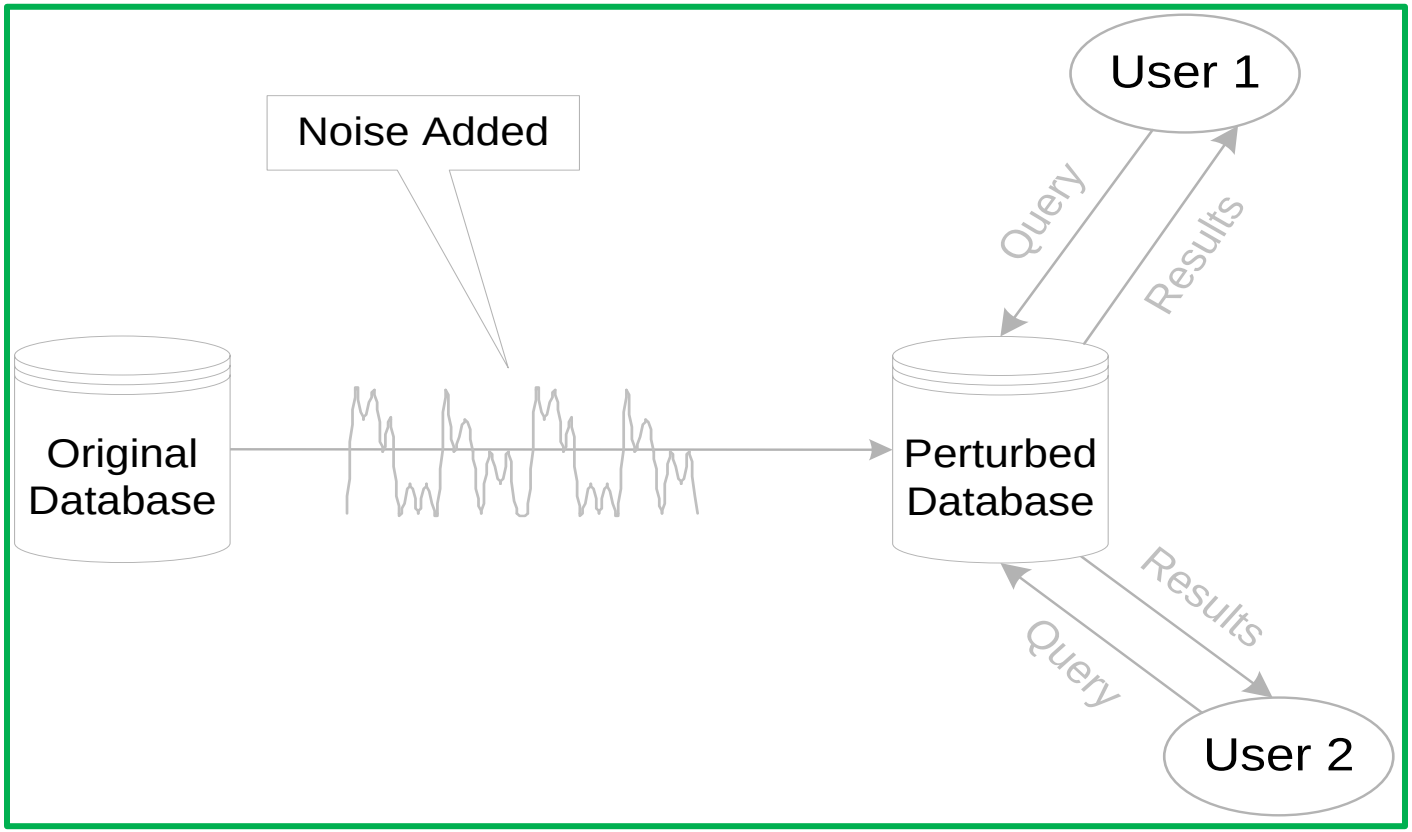


- Kỹ thuật gây nhiễu đầu ra



KỸ THUẬT GÂY NHIỄU DỮ LIỆU (Data Perturbation)

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



- 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,...,N) của một thuộc tính A_j bị thay thế bằng một giá trị gây nhiễu x'_{ij}

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

  - Vector $e = (x' - x) = (e_1,..., e_N)$ là một vector gây nhiễu ngẫu nhiên
  - $x = (x_{1j} ,..., x_{Nj})$, $x'=(x'_{1j} ,..., x'_{Nj})$ 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)
  - $e = (e_1, ..., e_N)$, mỗi thành phần e_i là các biến ngẫu nhiên, độc lập tuyến tính.

$$E(e_i) = 0,\ D(e_i) = \sigma^2$$

  - Các giá trị của mỗi thuộc tính A_j sẽ được cộng thêm một vector e ngẫu nhiên.
  - Xác suất lỗi trong một câu truy vấn vượt quá giá trị giới hạn **ε** cho trước là:
  - $P(|q'(C) - q(C)| >= |\varepsilon |X(C)| |) <= \sigma^2/(|X(C)|\varepsilon^2)$
  - 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 nhiều 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

- 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.

- Gây nhiễu dựa vào truy vấn
  - *Thống kê Sum:*

+ Xét thống kê $S = q(C) = Sum(C, Aj)$, n là số lu$\sum_{i=1}^{n} x'_{ij}$ các bản ghi tập truy vấn X($\overline{x_{C_j}}$

+ S' = với $x'_{ij} = f(x_{ij}) = x_{ij} + z_1 (x_{ij} -$)

+ z_2

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

- Gây nhiễu dựa vào truy vấn
  - Thống kê Count:

+ Gi$\sum_{j=3}^{n} z_3$ thống kê *Count(C) = m*

+ *m'* =

Với E(z_3) = 1 và Var(z_3) = a^2_1 /m,

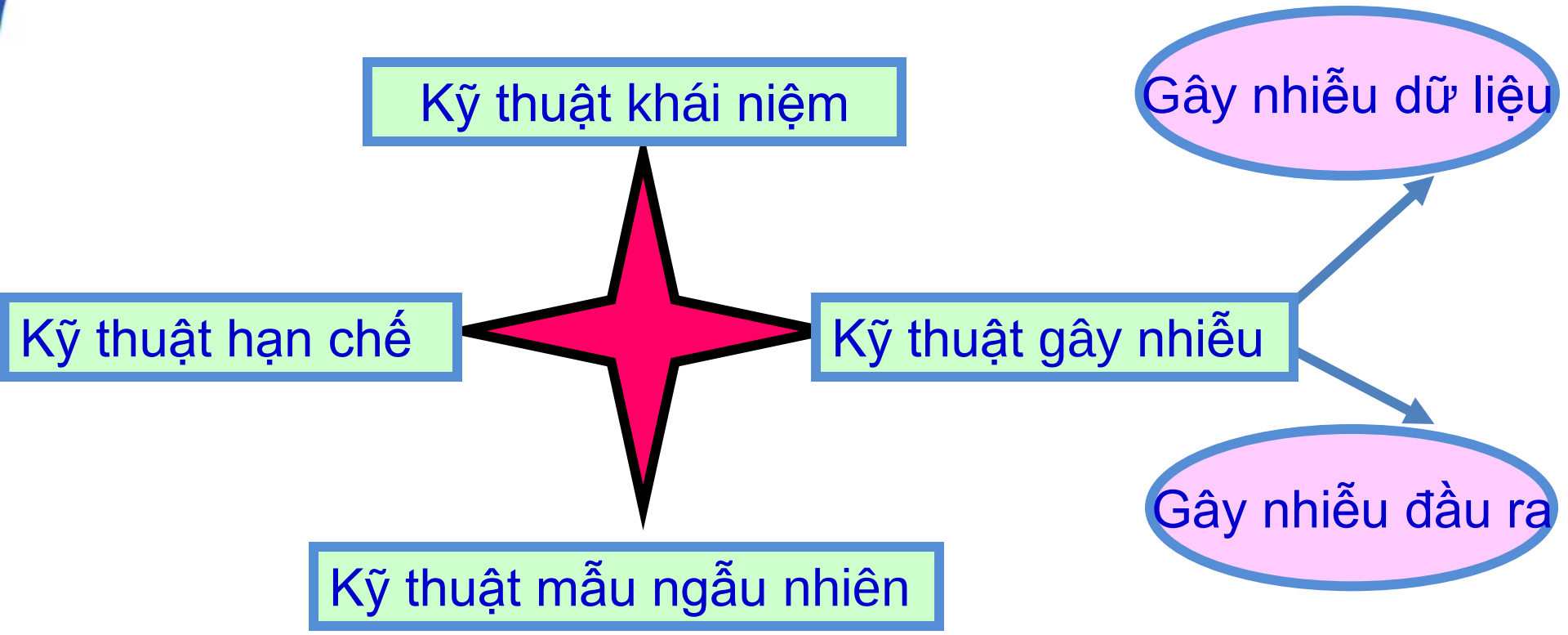
+ và z_3 được sinh ngẫu nhiên và độc lập với các bản ghi x_i trong X(C).

+ *E(m') = m* và *Var(m') = a^2_1*

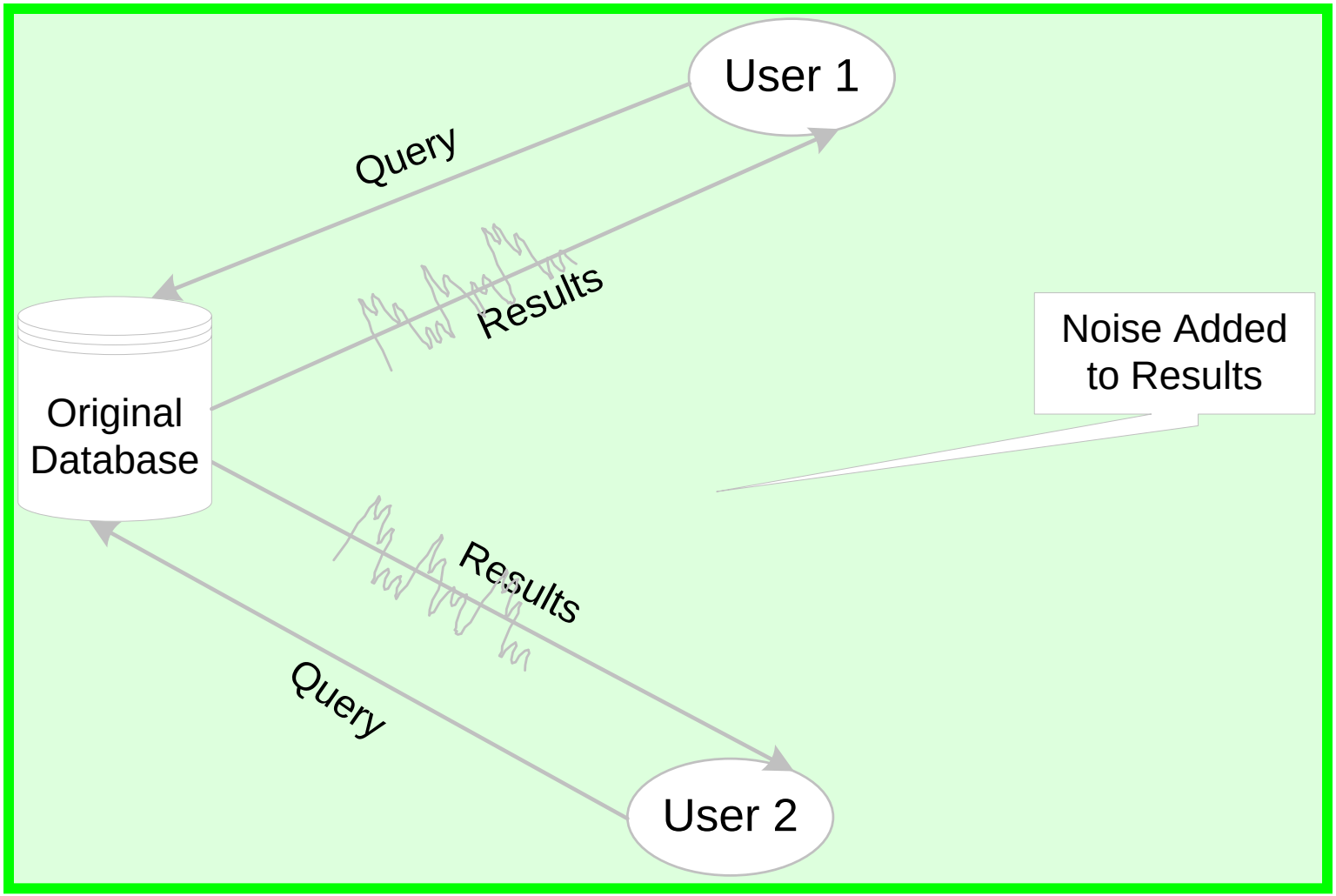
KỸ THUẬT GÂY NHIỄU DỮ LIỆU

- 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 gía trị nhiễu=> 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 mod b.
  - r(Q) = $\begin{cases} Q \ \ nêu\ d = 0 \\ Q - d \ \ nêu\ d < b' \\ Q + b - d \ \ nêu\ d \geq b' \end{cases}$

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 mod b.
  - r(Q) = $\begin{cases} Q \ \ nêu\ d = 0 \\ Q - d \ \ voi\ xác\ suât\ 1 - p \\ Q + b - d \ \ voi\ xác\ suât\ p \end{cases}$

Xác suất p = d/b

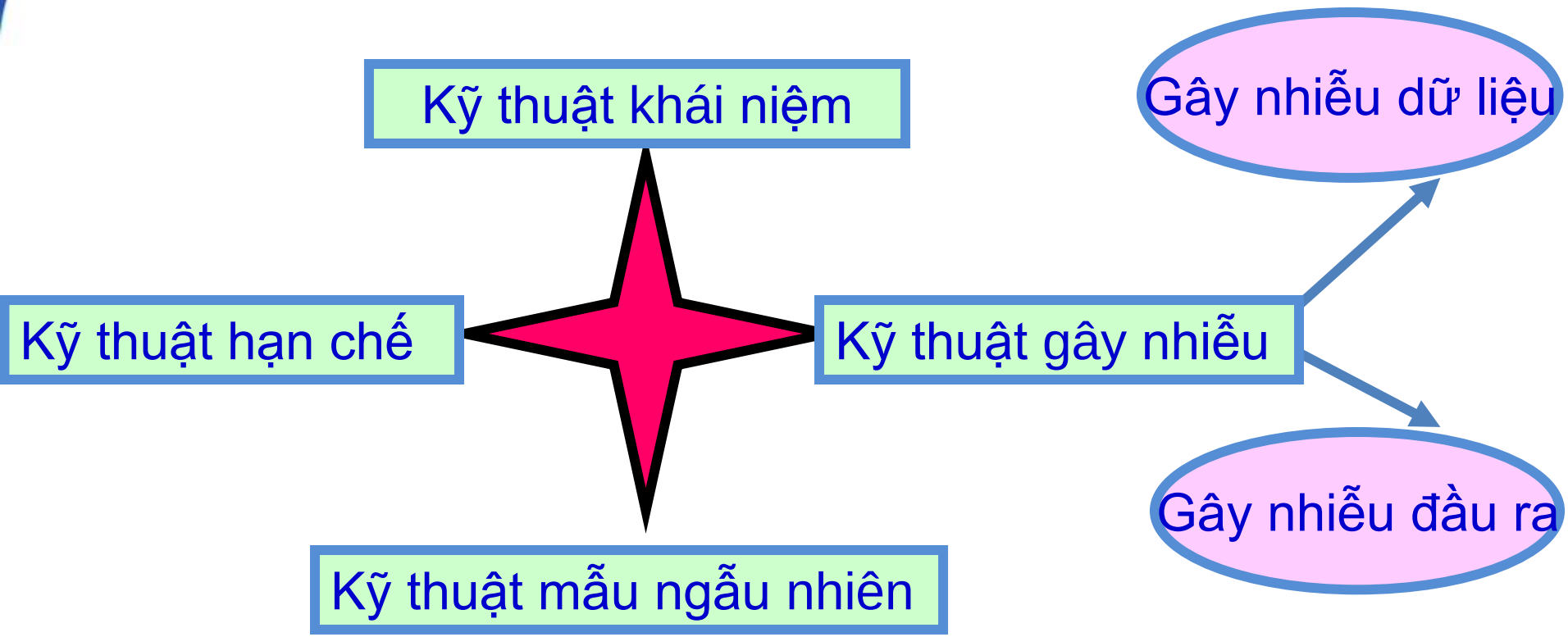
KỸ THUẬT GÂY NHIỄU ĐẦU RA

- 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.

NỘI DUNG

1. Kỹ thuật khái niệm
2. Kỹ thuật hạn chế
3. Kỹ thuật gây nhiễu
4. Kỹ thuật mẫu ngẫu nhiên

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.

KỸ THUẬT MẪU NGẪU NHIÊN

- 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.

NỘI DUNG

1. Kỹ thuật khái niệm
2. Kỹ thuật hạn chế
3. Kỹ thuật gây nhiễu
4. Kỹ thuật mẫu ngẫu nhiên

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 ngươì 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


QUESTION