COMPUTER SCIENCE
& ENGINEERING
TEXAS A&M UNIVERSITY

Robust De-anonymization of Large Datasets

Dr. Martin "Doc" Carlisle

Premise

- Can we extract personal info from anonymized database?

Netflix Prize Dataset

- Movie ratings of 500,000 Netflix subscribers
- “The **Netflix Prize** was an open competition for the best collaborative filtering algorithm to predict user ratings for films, based on previous ratings without any other information about the users or films, i.e. without the users or the films being identified except by numbers assigned for the contest.” (Wikipedia)

K-anonymity

- Publisher decides which attributes public/private
  - Public are “quasi-identifiers”
- Every quasi-identifier tuple appears in at least k records in anonymized DB

Wikipedia Example

Name	Age	Gender	State of domicile	Religion	Disease
Ramsha	30	Female	Tamil Nadu	Hindu	Cancer
Yadu	24	Female	Kerala	Hindu	Viral infection
Salima	28	Female	Tamil Nadu	Muslim	TB
Sunny	27	Male	Karnataka	Parsi	No illness
Joan	24	Female	Kerala	Christian	Heart-related
Bahuksana	23	Male	Karnataka	Buddhist	TB
Rambha	19	Male	Kerala	Hindu	Cancer
Kishor	29	Male	Karnataka	Hindu	Heart-related
Johnson	17	Male	Kerala	Christian	Heart-related
John	19	Male	Kerala	Christian	Viral infection

Wikipedia Example

- Suppress some fields (e.g. name, religion)
- Generalize some fields

Name	Age	Gender	State of domicile	Religion	Disease
*	20 < Age ≤ 30	Female	Tamil Nadu	*	Cancer
*	20 < Age ≤ 30	Female	Kerala	*	Viral infection
*	20 < Age ≤ 30	Female	Tamil Nadu	*	TB
*	20 < Age ≤ 30	Male	Karnataka	*	No illness
*	20 < Age ≤ 30	Female	Kerala	*	Heart-related
*	20 < Age ≤ 30	Male	Karnataka	*	TB
*	Age ≤ 20	Male	Kerala	*	Cancer
*	20 < Age ≤ 30	Male	Karnataka	*	Heart-related
*	Age ≤ 20	Male	Kerala	*	Heart-related
*	Age ≤ 20	Male	Kerala	*	Viral infection

2-anonymity
For Age,Gender,State

Model

- Consider database as NxM matrix
  - Rows are people
  - Columns are preferences
  - Very sparse! (what percentage of Netflix have you watched?)
    - Supp(r) is set of non-null attributes of a record

Similarity

- Measure similarity using generalized cosine measure
  - Sim on two attributes maps to [0,1]
  - E.g. a binary if ratings/dates are within threshold between subscribers

$$\text{Sim}(r_1, r_2) = \frac{\sum \text{Sim}(r_{1i}, r_{2i})}{|\text{supp}(r_1) \cup \text{supp}(r_2)|}$$

Sparsity definition

- Netflix Prize database vast majority has no record with similarity over 0.5, even if consider only movies rated (not dates or ratings)
  - (0.5,0) sparse

Definition 1 (Sparsity) *A database D is (ϵ, δ)-sparse w.r.t. the similarity measure* Sim *if*

$$\Pr_r[\mathit{Sim}(r, r') > \epsilon \; \forall r' \neq r] \leq \delta$$

Similarity

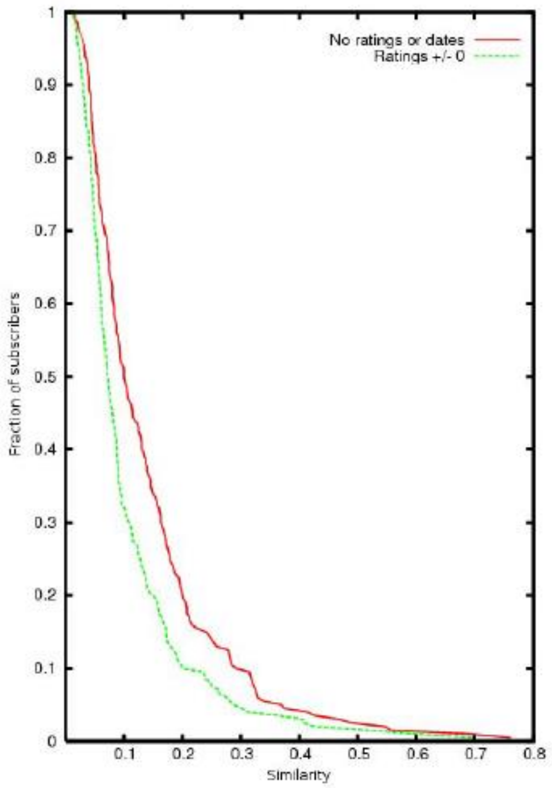

Figure 1. X-axis (x) is the similarity to the "neighbor" with the highest similarity score; Y-axis is the fraction of subscribers whose nearest-neighbor similarity is at least x.

Adversary model

- Has auxiliary information (background info related to record)
  - Aux(r)
- Adversary wants to reconstruct values of entire record r given auxiliary info and anonymized sample

Deanonymized? (I)

- Amplification of background knowledge
- Uses Aux(r) close to r on subset of attributes to find r' close to r on all

Definition 2 *A database D can be (θ, ω)-deanonymized w.r.t. auxiliary information* Aux *if there exists an algorithm A which, on inputs D and* $\mathsf{Aux}(r)$ *where $r \leftarrow D$ outputs r' such that*

$$\Pr[\mathit{Sim}(r, r') \geq \theta] \geq \omega$$

Deanonymized? (II)

- Extended to a subset

Definition 3 (De-anonymization) *An arbitrary subset $\hat{D}$ of a database D can be (θ, ω)-deanonymized w.r.t. auxiliary information* Aux *if there exists an algorithm A which, on inputs $\hat{D}$ and* $\mathsf{Aux}(r)$ *where $r \leftarrow D$*

- *If $r \in \hat{D}$, outputs r' s.t.* $\Pr[\mathsf{Sim}(r, r') \geq \theta] \geq \omega$
- *if $r \notin \hat{D}$, outputs $\perp$ with probability at least ω*

Entropic de-anonymization

- Measures entropy of candidate set of records similar to target record

Definition 4 (Entropic de-anonymization) *A database D can be (θ, H)-deanonymized w.r.t. auxiliary information* Aux *if there exists an algorithm A which, on inputs D and* $\mathsf{Aux}(r)$ *where $r \leftarrow D$ outputs a set of candidate records D' and probability distribution Π such that*

$$E[\mathit{min}_{r' \in D', \mathit{Sim}(r,r') \geq \theta} H_S(\Pi, r')] \leq H$$

De-anonymization Algorithm

- Inputs, database sample + auxiliary information for a record
- Output: record in sample, or set with probabilities

1. Compute score(aux, r’) for each r’ in sample
2. Apply matching criteria
3. Output record or probability distribution for records

Algorithm Scoreboard

- $\mathsf{Score}(\mathsf{aux}, r') = \mathsf{min}_{i \in \mathsf{supp}(\mathsf{aux})} \mathsf{Sim}(\mathsf{aux}_i, r'_i)$, *i.e.*, the score of a candidate record is determined by the least similar attribute between it and the adversary's auxiliary information.
- The matching set $D' = \{r' \in \hat{D} : \mathsf{Score}(\mathsf{aux}, r') > \alpha\}$ for some fixed constant α. The matching criterion is that D' be nonempty.
- Probability distribution is uniform on D'.

Netflix Prize Dataset

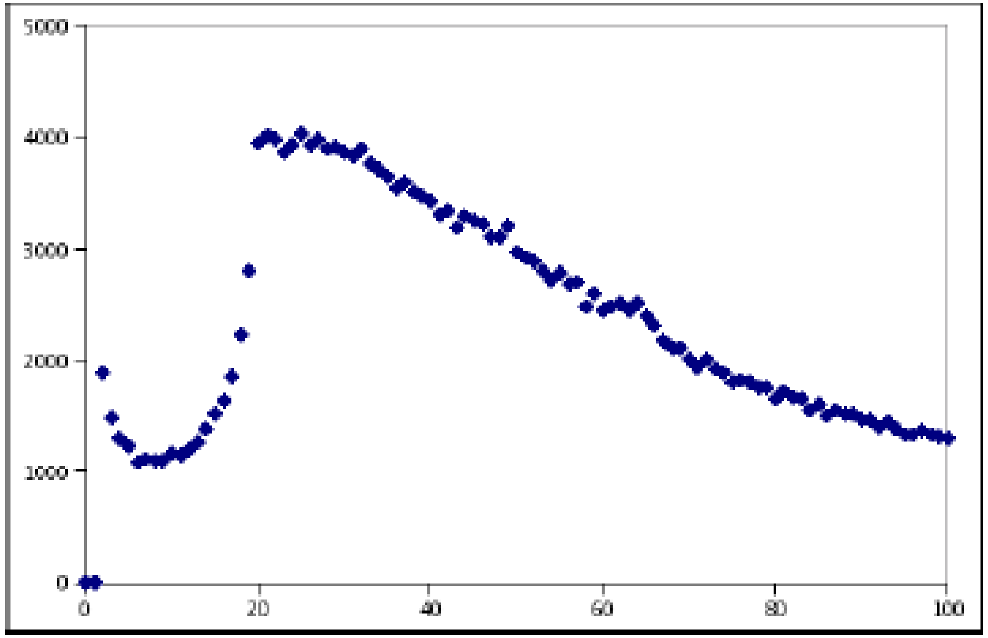

Figure 2. For each $X \leq 100$, the number of subscribers with X ratings in the released dataset.

What if I know some approximate dates with ratings?

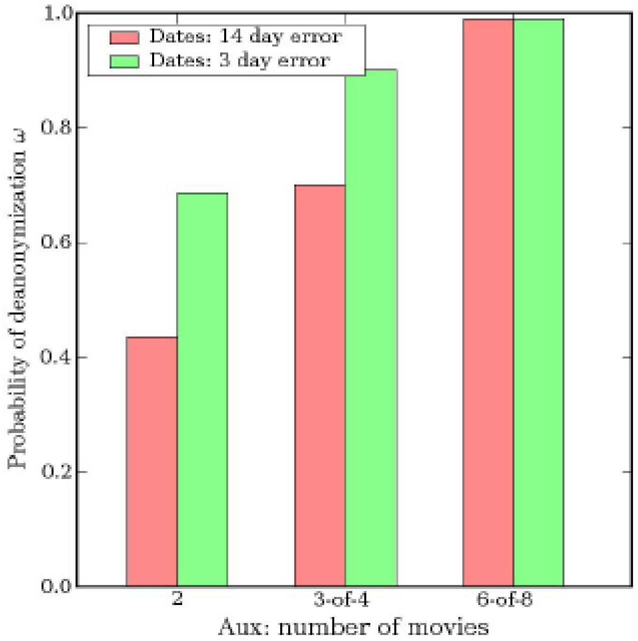

Figure 4. Adversary knows exact ratings and approximate dates.

Maybe you are not in there?

The red (darker) bars represent the probability of correctly detecting that the record is not in the sample

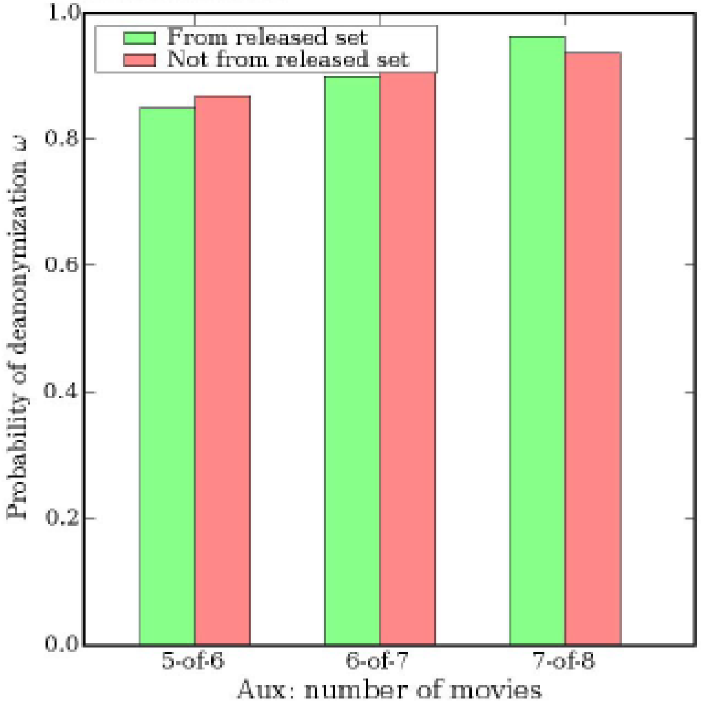

Figure 5. Same parameters as Fig. 4, but the adversary must also detect when the target record is not in the sample.

Ratings but no dates

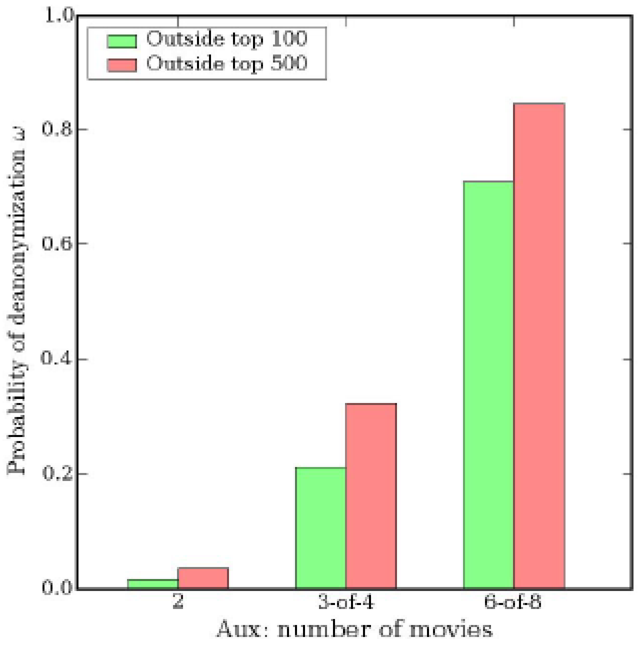

Figure 8. Adversary knows exact ratings but does not know dates at all.

Less popular movies

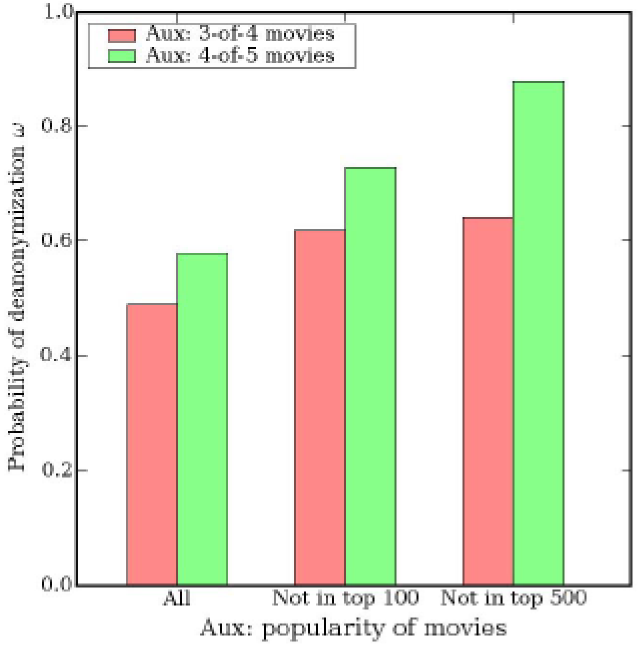

Figure 9. Effect of knowing less popular movies rated by victim. Adversary knows approximate ratings (± 1) and dates (14-day error).