The Netflix Paper

Daniel King
cs252r Fall ‘13

"… we succesfully identified the Netflix records of known users, uncovering their apparent political preferences …"

–Narayanan & Shmatikov

The Netflix Prize

- 1 million USD prize
- Improve Netflix's recommendation service
- Publicly released 100,480,507 movie ratings
- 480,189 Netflix subscribers
- December 1999 to December 2005

Netflix Prize FAQ

- Q: Is there any customer information in the dataset that should be kept private?

Netflix Prize FAQ

- A: “No, all customer identifying information has been removed; all that remains are ratings and dates …”

Netflix Prize FAQ

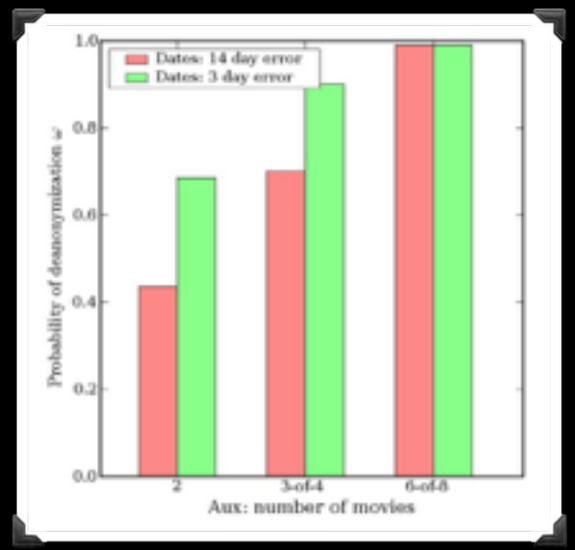
- A: “Even if, for example, you knew all your own ratings and their dates you probably couldn’t identify them reliability”

Is this true?

No.

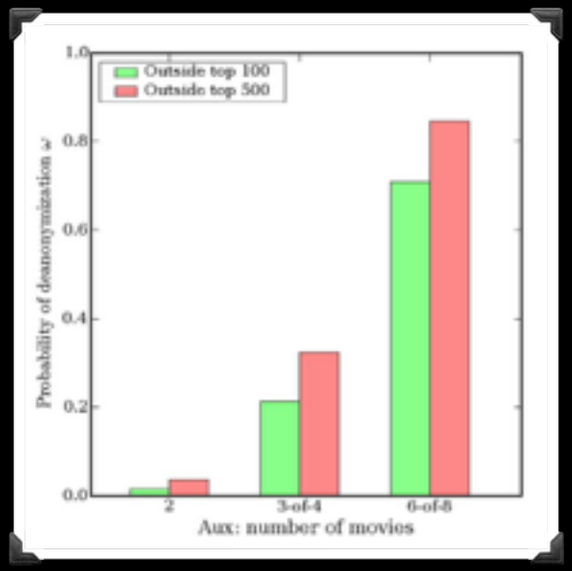
De-anonymizing

- 6-of-8 means the adversary, a priori, knows eight movie ratings of the target
- Six of the ratings are correct
- Two of the ratings are wrong
- Knowledge of six ratings slashes uncertainty to less than a bit!



De-anonymizing

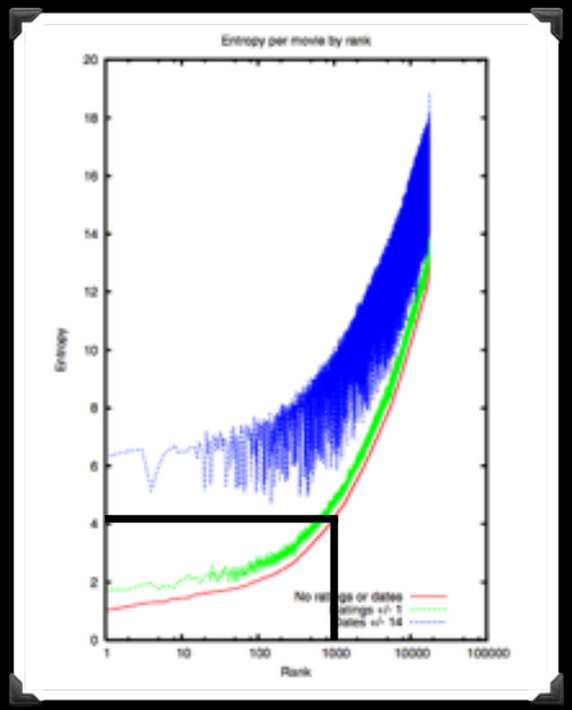
- Adversary has no information about date watched
- Six correct ratings still reveals user with 70 to 80% probability



What if the adversary doesn't know timing information?

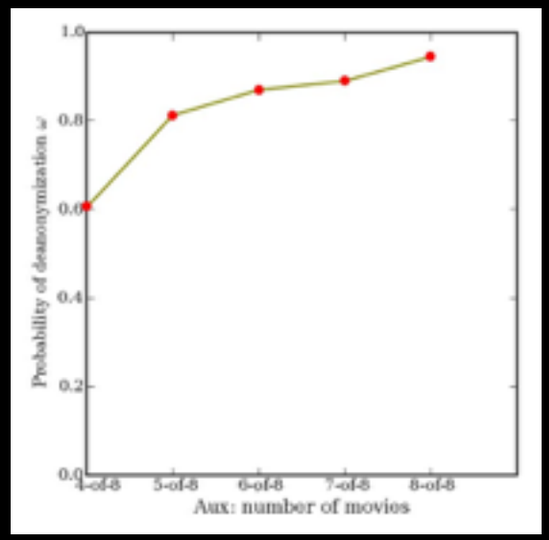
De-anonymizing

- Simply *watching* a top 1000 movie reveals 4 bits of information
- The dataset in total has 19 bits of entropy



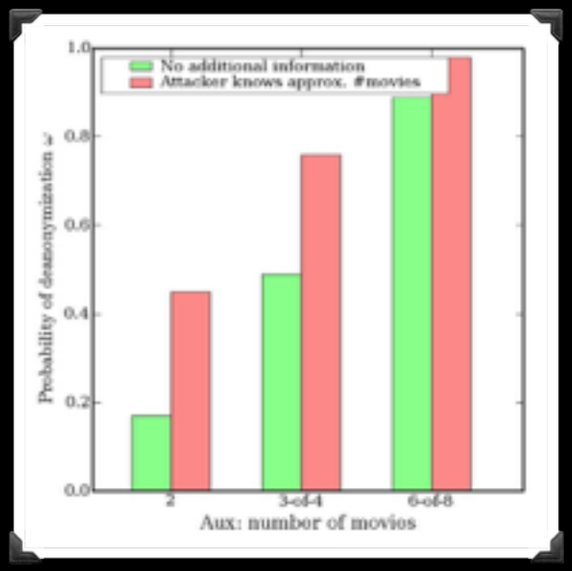
De-anonymizing

- Number of correct ratings can fall to 5 of 8 and still provide 80% probability of de-anonymization



De-anonymizing

- Adversary knows number of movies *with a 50% error*
- Knows dates within 14-days
- Knows ratings ±1



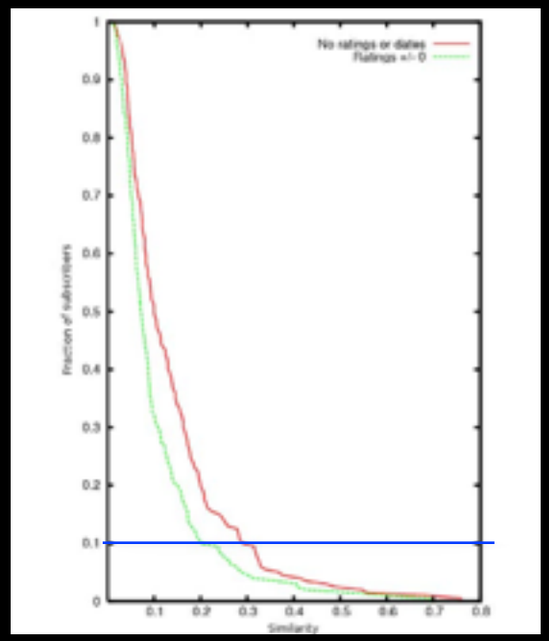
Discriminating users by number of movies watched can double probability of de-anonymization

Exploitable Features

- Sparsity
- Auxiliary Data

Sparsity

- Similarity is a measure of how many ratings are equivalent
- 90% of Netflix users stand out significantly from their peers
- Almost uniquely represented by vector of ratings

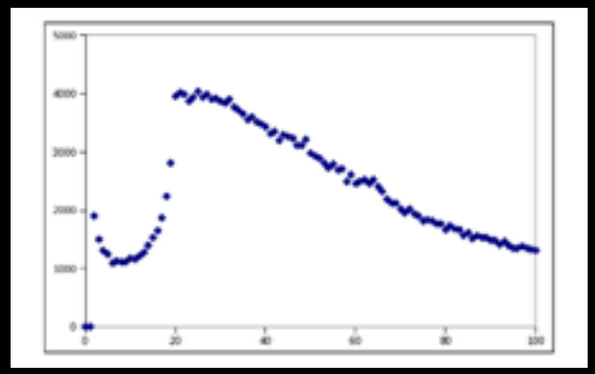


High dimensional data about humans is sparse

We’re unique individuals whose interests cause this

Sparsity

- x-axis is number of movie ratings
- y-axis is number of users who have rated that number of movies
- Individuals still stand out among *thousands* of peers.



Auxiliary Data

- Public IMDb ratings
- Private moving watching habits become public
- Forward Secrecy

The adversary, say an unsavory hiring manager, finds a user's public IMDb record, uses this to find out that the user has watched movies the manager finds disagreeable.

If Netflix dataset is public, the users can never again publicly acknowledge watching or rating any movies in that set for fear of being identified. Forward Secrecy

Formalism

- Datasets are matrices, often sparse
- Compare the “similarity” of datasets
- Adversaries are represented by their “auxiliary” data
- Privacy breach is defined probabilistically

Datasets are modeled as matrices, a notion of similarity between user rows is given, an adversary is modelled as a an Auxiliary function, a formal notion of privacy breach is given in terms of the probability that a row

Dataset

- a matrix
- rows are users
- columns are attributes

	Pi	Love Actually	Inception	…
Dan	5	4	0	…
…	…	…	…	…

The dataset is modeled as a matrix where each row is associated with a user and columns are associated with attributes. For example, I might be row 35 and I'd have a five in Pi, a four in Love Actually, and a never watched in Inception.

Similarity

- Codomain of similarity is [0,1]
- Domain of similarity is either rows or attribute values
- Similarity on attribute values is the “indicator” function

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

In order to define privacy breach we need to understand what it means to correctly guess a user

indicator function is 1 for equivalence 0 for non-equivalence

Similarity

- |…| is the length of the row
- supp(…) elides zero elements
- Sim(r1,r1) = 1

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

In order to define privacy breach we need to understand what it means to correctly guess a user

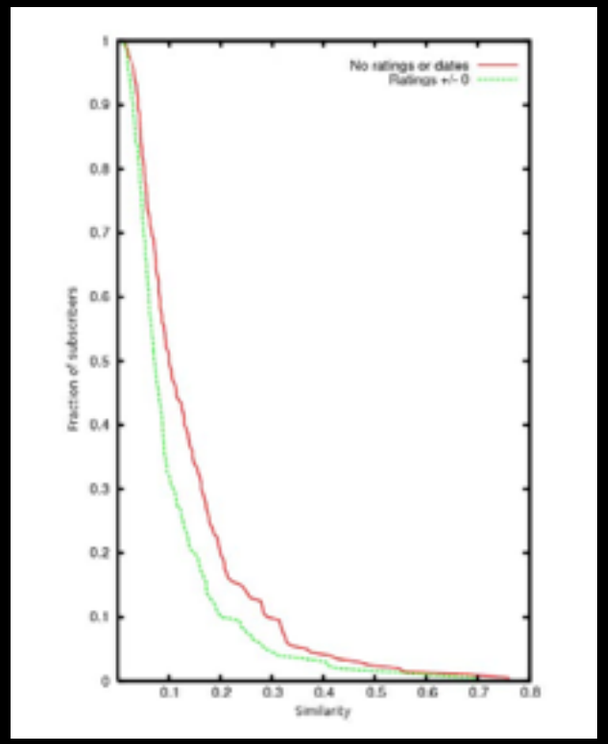
indicator function is 1 for equivalence 0 for non-equivalence

Sparsity

- (0.5,1e-5)-sparsity means for a dataset of 500,000 users, five users share more than half their movie ratings in common
- In the Netflix dataset, no records have in common more than half of their movie ratings

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

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



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

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

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

In light of that statement this graph is a big confusing

Adversary

- Accesses some auxiliary data
- Possibly modified or incorrect

The IMDb data is assumed to come from the same underlying dataset, how a group of people think about a set of movies

Adversary

- Sanitization of dataset can be modeled instead in auxiliary data
- Aux : RowOfAttributes -> RowOfAttributes

Discrepancies between IMDb ratings and Netflix ratings are modeled in the Aux function, the output row could contain many zero, “unwatched” entries.

De-anonymization

- The adversary is given:
  - the image of Aux on the hidden dataset
  - the dataset

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* **Aux**(r) *where $r \leftarrow D$ outputs r' such that*

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

In this case, Aux models sanitized data and the adversary has access to the full dataset and needs to identify sanitized records with real records

De-anonymization

- Adversary can report that the record doesn't appear in the 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* **Aux**(r) *where $r \leftarrow D$*

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

This is a more realistic model where the adversary only gets access to a subset of the data and is permitted to state that the user doesn't paper in the subset.

Entropic De-anonymization

- Use the minimum Shannon entropy
- Not quite the same as min entropy

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* **Aux**(r) *where $r \leftarrow D$ outputs a set of candidate records D' and probability distribution Π such that*

$$E[\min_{r' \in D', \mathbf{Sim}(r,r') > \theta} H_{\Sigma}(\Pi, r')] \leq H'$$

Let's the algorithm output a probability distribution over possible matches

Entropic De-anonymization

- Use the minimum Shannon entropy
- Not quite the same as min entropy

$$H_S = \min\left[-\mathbb{E}[log(\Pi_i)]\right]$$

$$H_\infty = \min\left[-log(\Pi_i)\right]$$

Algorithm Scoreboard

Algorithm Scoreboard. The following simple instantiation of the above template is sufficiently tractable to be formally analyzed in the rest of this section.

- $\mathsf{Score}(\mathsf{aux}, r') = \min_{i \in \mathsf{supp}(\mathsf{aux})} \mathsf{Sim}(\mathsf{aux}_i, r'_i)$, *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'.

Fails if the auxiliary information is wrong on any rating

Algorithm Scoreboard-RH

- $\mathsf{Score}(\mathsf{aux}, r') = \sum_{i \in \mathsf{supp}(\mathsf{aux})} \mathsf{wt}(i)\mathsf{Sim}(\mathsf{aux}_i, r'_i)$ where $\mathsf{wt}(i) = \frac{1}{\log |\mathsf{supp}(i)|}$.[3]
- If a "best guess" is required, compute $\mathsf{max} = \mathsf{max}(S), \mathsf{max}_2 = \mathsf{max}_2(S)$ and $\sigma = \sigma(S)$ where $S = \{\mathsf{Score}(\mathsf{aux}, r') : r' \in \hat{D}\}$, *i.e.*, the highest and second-highest scores and the standard deviation of the scores. If $\frac{\mathsf{max} - \mathsf{max}_2}{\sigma} < \phi$, where ϕ is a fixed parameter called the *eccentricity*, then there is no match; otherwise, the matching set consists of the record with the highest score.[4]
- If entropic de-anonymization is required, output distribution $\Pi(r') = c \cdot e^{\frac{\mathsf{Score}(\mathsf{aux}, r')}{\sigma}}$ for each r', where c is a constant that makes the distribution sum up to 1. This weighs each matching record in inverse proportion to the likelihood that the match in question is a statistical fluke.

- supp(attribute) is the number of users with non-zero values for that attribute
- more heavily weight rare attributes
- if a best guess is required (the non-entropy definitions) output the highest if it “stands out”
- otherwise there is no match

Results

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

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 Aux(r) where $r \leftarrow D$ outputs r' such that*

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

Algorithm Scoreboard. The following simple instantiation of the above template is sufficiently tractable to be formally analyzed in the rest of this section.

- $\mathsf{Score}(\mathsf{aux}, r') = \min_{i \in \mathsf{supp}(\mathsf{aux})} \mathsf{Sim}(\mathsf{aux}_i, r'_i)$, *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'.

Let aux be the auxiliary information about some record r; aux consists of m (non-null) attribute values, which are close to the corresponding values of attributes in r, that is, $|\mathsf{aux}| = m$ and $\mathsf{Sim}(\mathsf{aux}_i, r_i) \geq 1 - \epsilon \; \forall i \in \mathsf{supp}(\mathsf{aux})$, where aux_i (respectively, r_i) is the ith attribute of aux (respectively, r).

Theorem 1 *Let $0 < \epsilon, \delta < 1$ and let D be the database. Let Aux be such that aux $=$ Aux(r) consists of at least $m \geq \frac{\log N - \log \epsilon}{-\log(1-\delta)}$ randomly selected attribute values of the target record r, where $\forall i \in \mathit{supp}(\mathit{aux})$, $\mathit{Sim}(\mathit{aux}_i, r_i) \geq 1 - \epsilon$. Then D can be $(1 - \epsilon - \delta, 1 - \epsilon)$-deanonymized w.r.t. Aux.*

Use algorithm with alpha = 1-epsilon

Results

Theorem 1 *Let $0 < \epsilon, \delta < 1$ and let D be the database. Let Aux be such that $aux = Aux(r)$ consists of at least $m \geq \frac{\log N - \log \epsilon}{-\log(1-\delta)}$ randomly selected attribute values of the target record r, where $\forall i \in supp(aux)$, $Sim(aux_i, r_i) \geq 1-\epsilon$. Then D can be $(1-\epsilon-\delta, 1-\epsilon)$-deanonymized w.r.t. Aux.*

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

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 $Aux(r)$ where $r \leftarrow D$ outputs r' such that*

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

Algorithm Scoreboard. The following simple instantiation of the above template is sufficiently tractable to be formally analyzed in the rest of this section.

- $\mathsf{Score}(\mathsf{aux}, r') = \min_{i \in \mathsf{supp}(\mathsf{aux})} \mathsf{Sim}(\mathsf{aux}_i, r'_i)$, *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'.

Lemma 1 *If r' is a false match, then $\Pr_{i \in \mathsf{supp}(r)}[\mathsf{Sim}(r_i, r'_i) \geq 1-\epsilon] < 1-\delta$*

Lemma 1 holds, because the contrary implies $\mathsf{Sim}(r, r') \geq (1-\epsilon)(1-\delta) \geq (1-\epsilon-\delta)$, contradicting the assumption that r' is a false match. Therefore, the probability that the false match r' belongs to the matching set is at most $(1-\delta)^m$. By a union bound, the probability that the matching set contains even a single false match is at most $N(1-\delta)^m$. If $m = \frac{\log \frac{N}{\epsilon}}{\log \frac{1}{1-\delta}}$, then the probability that the matching set contains any false matches is no more than ϵ.

Therefore, with probability $1-\epsilon$, there are no false matches. Thus for every record r' in the matching set, $\mathsf{Sim}(r, r') \geq 1-\epsilon-\delta$, *i.e.*, any r' must be similar to the true record r. To complete the proof, observe that the matching set contains at least one record, r itself.

When δ is small, $m = \frac{\log N - \log \epsilon}{\delta}$. This depends logarithmically on ϵ and linearly on δ: the chance that the algorithm fails completely is very small even if attribute-wise accuracy is not very high. Also note that the matching set need not be small. Even if the algorithm returns many records, with high probability they are *all* similar to the target record r, and thus any one of them can be used to learn the unknown attributes of r.

Use algorithm with alpha = 1-epsilon

Results

- A (1-ε-δ, ε)-sparse dataset can be (1, 1-ε)-deanonymized

Theorem 2 *Let ϵ, δ, and aux be as in Theorem 1. If the database D is $(1-\epsilon-\delta, \epsilon)$-sparse, then D can be $(1, 1-\epsilon)$-deanonymized.* □

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

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

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 $Aux(r)$ where $r \leftarrow D$ outputs r' such that*

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