# On the Unicity of Smartphone Applications

**Jagdish Prasad Achara**[Speaker]
**Gergely Acs**
**Claude Castelluccia**

**INRIA, France**

# Previous studies

- Visited locations are unique
  - *Unique in the Crowd: The privacy bounds of human mobility,*

    **Nature Scientific Report, 2013**

- Credit card metadata is unique
  - *On the Re-identifiability of credit card metadata,*

    **Science, 2015**

# Why we are interested in Unicity?

# Release of pseudo-anonymized dataset

| Rec # | Data |
|---|---|
| **1** | {Item 2, Item 3, Item 8, …, Item M} |
| **2** | {Item 1, Item 13, …, Item N} |
| … | … |

What is the risk of Re-identification of users?

# Measuring Re-identification risk

**Dataset D1**

| Rec # | Data |
|---|---|
| **1** | {Item 2, Item 3, Item 8, Item 6} |
| **2** | {Item 1, Item 4} |
| **3** | {Item 2, Item 4, Item 6} |

Unicity(2 items in D1) = 8/9

**Dataset D2**

| Rec # | Data |
|---|---|
| **1** | {Item 2, Item 3, Item 8, Item 6} |
| **2** | {Item 2, Item 3, Item 7, Item 6} |
| **3** | {Item 2, Item 3, Item 1} |

Unicity(2 items in D2) = 6/9

Re-identification risk ≈ Unicity of K-items

# How to measure Unicity of K-items?

# Unicity Measurement

| Rec # | Data |
|---|---|
| **1** | {Item 2, Item 3, Item 8, …, Item M} |
| **2** | {Item 1, Item 13, …, Item N} |
| … | … |

$$Unicity\ of\ K\text{-}items = \frac{\#\ of\ K\text{-}items\ that\ appear\ once}{total\ \#\ of\ K\text{-}items\ in\ the\ dataset}$$

Prohibitively expensive to calculate,

**sampling as rescue.**

# How to sample?

- Naïve technique:
  1. Randomly select a user having ≥ K items
  2. Randomly select K-items from that user

| Rec # | Data |
|---|---|
| 1 | {Item 4, Item 5, Item 8, Item 11} |
| 2 | {Item 8, Item 6, Item 11} |
| 3 | {Item 5, Item 7} |
| 4 | {Item 1, Item 11, Item 8} |

# How to sample?

- Naïve technique:
  1. Randomly select a user having ≥ K items
  2. Randomly select K-items from that user

- **Also state-of-the-art technique**
  - *Unique in the Crowd: The privacy bounds of human mobility,*

    **Nature Scientific Report, 2013**
  - *On the Re-identifiability of credit card metadata,* **Science, 2015**

# How to sample?

- Naïve technique:
  1. Randomly select a user having ≥ K items
  2. Randomly select K-items from that user

**Biased towards selecting more popular items!**

# How to sample?

State-of-the-art technique is biased!

- **Example:**

| Rec # | Data |
|---|---|
| 1 | {Item 4, Item 5, Item 8, Item 11} |
| 2 | {Item 8, Item 6, Item 11, Item 13} |
| 3 | {Item 5, Item 7, Item 2, Item 3} |
| 4 | {Item 1, Item 11, Item 8, Item 12} |

*Probability(Items 8, 11 are selected) = ¾*⅙*

*Probability(Items 5, 7 are selected) = ¼*⅙*

# How to uniformly sample?

- Naïve approaches are inappropriate
  - Rejection sampling
  - Enumerate all possible combinations of K-apps and find their support in the dataset
- Worst-case complexity is exponential in K

# Proposed uniform sampling technique

- Based on Metropolis-Hastings algo
  - A Markov Chain Monte Carlo (MCMC) method
- We construct an ergodic Markov chain ($\mathcal{M}$)
  - Such that its stationary distribution is uniform
- Every possible K-items represent a state of $\mathcal{M}$
  - Simulate $\mathcal{M}$ until it gets close to uniform

# Proposed uniform sampling technique

- MCMC sampling algorithm ($\mathcal{M}$):

Start with any existing K-items in the dataset.
REPEAT
1. PROPOSAL:
    1.1 sample a user uniformly at random
    1.2 select K-apps C from this user also uniformly at random
2. PROBABILISTIC ACCEPTANCE:
    2.1 accept it (i.e., S=C) with a probability, which is
        min(1, Pr["S is proposed"]/Pr["C is proposed"])
UNTIL Convergence

# Proposed uniform sampling technique

- Mixing time: roughly the order of dataset size
  - Real convergence much smaller for our dataset

- Overall Worst-case complexity $O(K|D|/H)$
  - $|D|$ is dataset size
  - H is the unicity of K-apps from the largest record of D

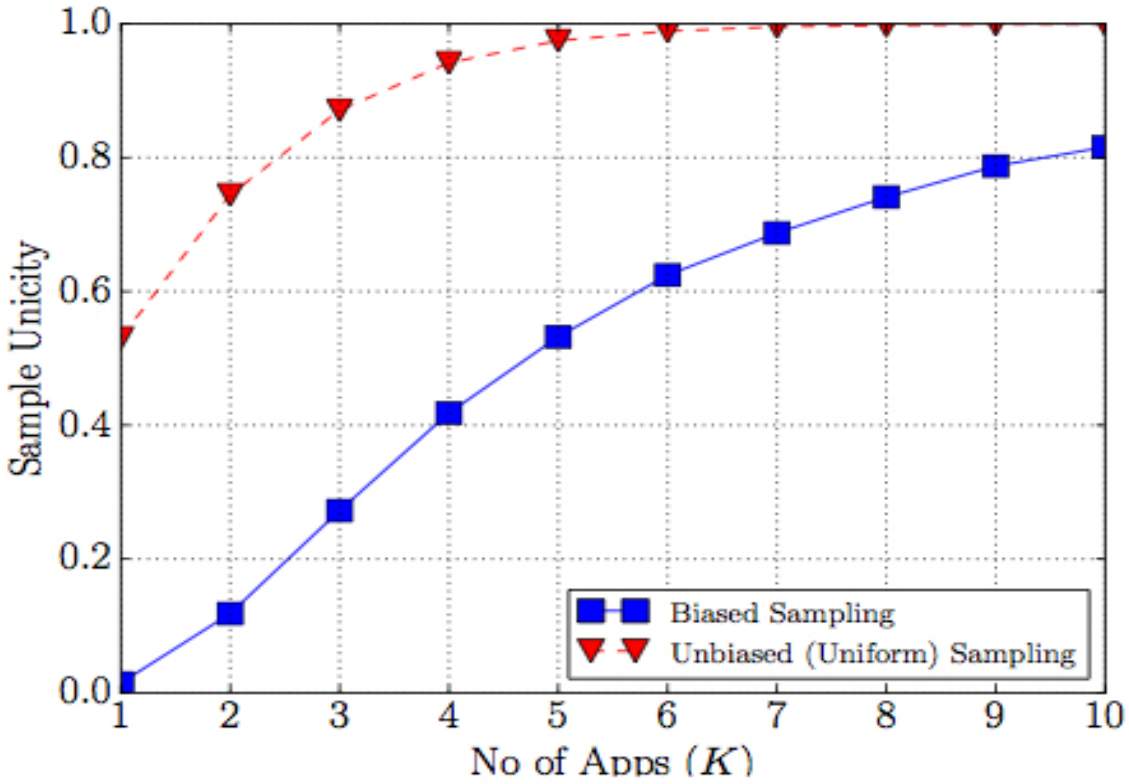
# Unicity of K-Apps

# Dataset

- Comes from Carat research project [1]
- Contains list of installed Apps of users
- 54,893 Android users
- 92,210 apps
- Collected over a period of 7 month

[1] ***http://carat.cs.helsinki.fi***

# Results




- *Unique in the Crowd: The privacy bounds of human mobility,* **Nature Scientific Report, 2013**
- *On the Re-identifiability of credit card metadata,* **Science, 2015**

# Parties knowing Installed Apps

- AppStore owners
  - Know all installed Apps
- Installed Apps themselves
  - May know all or a subset
- Included libraries (Ad, Analytics etc.)
  - May know all or a subset
- Friends/relatives
  - May know a subset

# Installed Apps are Quite Revealing




**Reference:**

- *Predicting User Traits from a Snapshot of Apps Installed on a Smartphone,* **ACM SIGMOBILE Mobile Computing and Communications Review, 2014**

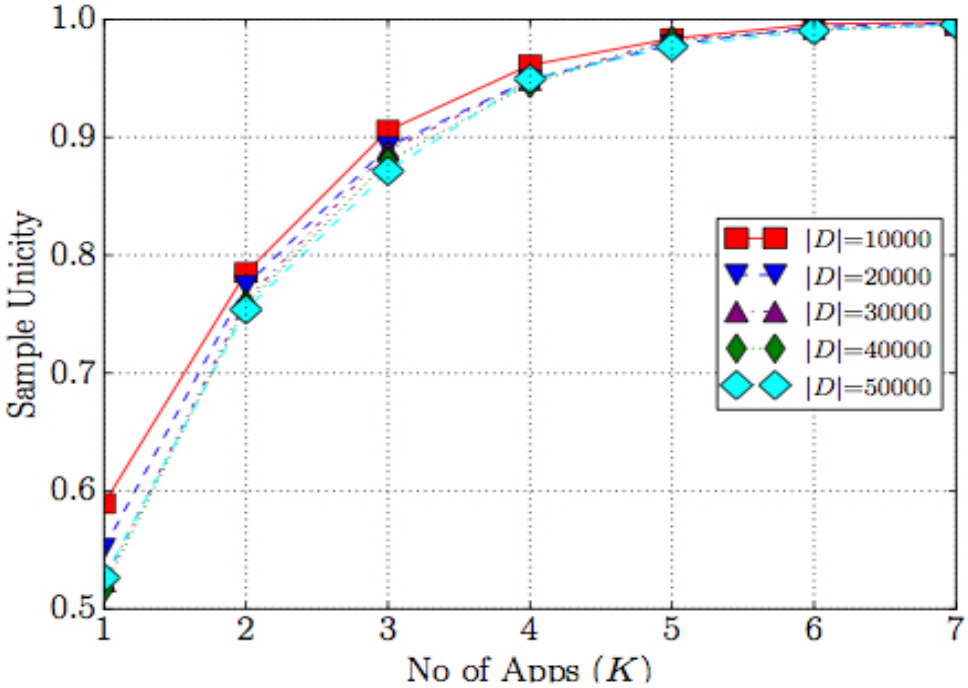
# Unicity generalization

# Regression analysis

- Created datasets of varying sizes and computed the sample Unicity

- Assumed Unicity a proper function of dataset size

# Model selection for regression analysis




$$f(x) = ae^{-b\sqrt{x}} + c$$

where x is the dataset size and f(x) is unicity value

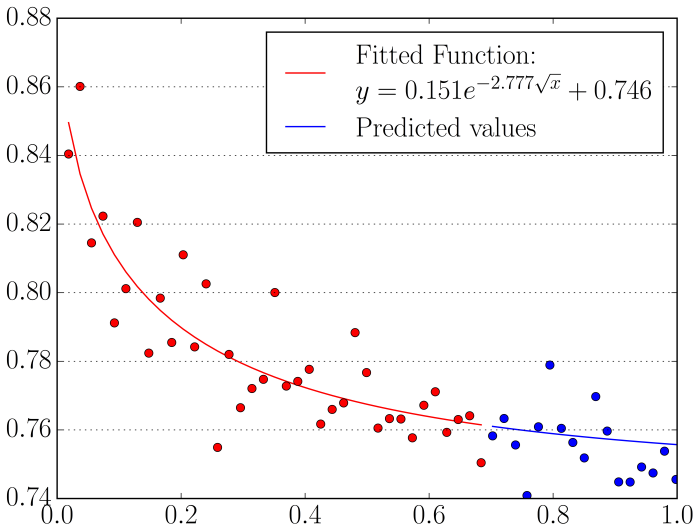
# Non-linear regression

- Data is divided into 50 parts (size 1K to 50K)
  - For training: first 70%
  - For testing: last 30%

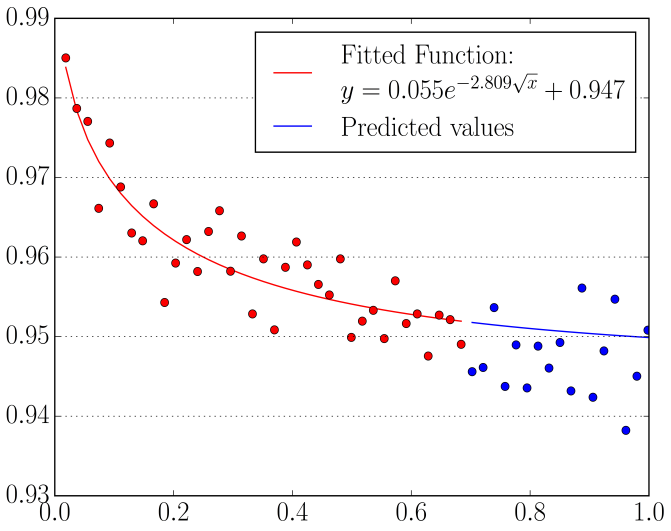
- Training data to learn a, b and c in the model

$$f(x) = ae^{-b\sqrt{x}} + c$$

# Unicity Generalization: App Dataset




(b) $K = 2$, $\delta = 0.007$




(d) $K = 4$, $\delta = 0.005$

δ is the avg. absolute error

# Conclusions

- Proposed a method for uniform sampling of K-items
- Addressed how Unicity would vary with dataset size
- Installed Apps are quite unique

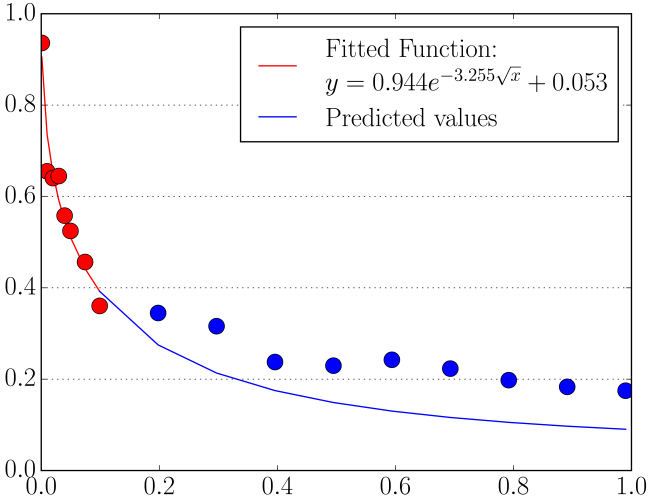
# Recommendations

- Access to the list of installed apps should be protected on Android
  - This is already the case with iOS 9
- Extreme care must be taken before releasing pseudo-anonymized installed Apps dataset
  - Adversaries do exist

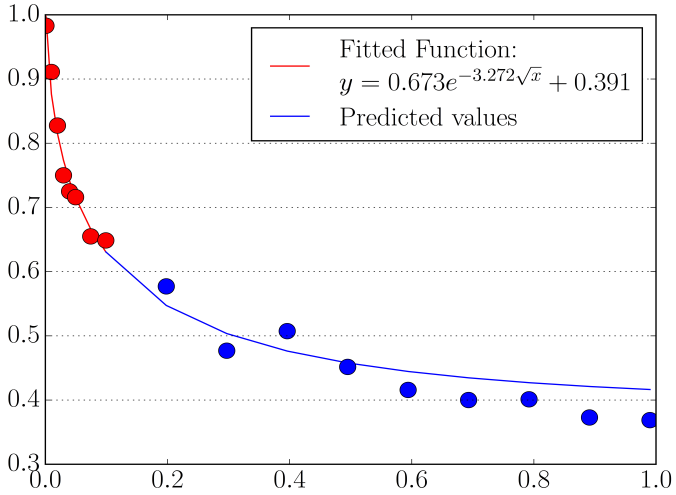
Thanks for your attention!
Questions?

# Backup slides

# Unicity Generalization: CDR data




(f) $K = 3$, trained with max $|D| = 100000$, $\delta = 0.089$




(c) $K = 4$, trained with max $|D| = 100000$, $\delta = 0.031$

δ is average absolute error