k-Anonymity and Other Cluster-Based Methods

Vitaly Shmatikov

Reading Assignment

- Li, Li, Venkatasubramanian. "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity" (ICDE 2007).

Background

- Large amount of person-specific data has been collected in recent years
  - Both by governments and by private entities
- Data and knowledge extracted by data mining techniques represent a key asset to the society
  - Analyzing trends and patterns.
  - Formulating public policies
- Laws and regulations require that some collected data must be made public
  - For example, Census data

Public Data Conundrum

- Health-care datasets
  - Clinical studies, hospital discharge databases ...
- Genetic datasets
  - $1000 genome, HapMap, deCode ...
- Demographic datasets
  - U.S. Census Bureau, sociology studies ...
- Search logs, recommender systems, social networks, blogs ...
  - AOL search data, social networks of blogging sites, Netflix movie ratings, Amazon ...

What About Privacy?

- First thought: anonymize the data
- How?
- Remove "personally identifying information" (PII)
  - Name, Social Security number, phone number, email, address... what else?
  - Anything that identifies the person directly
- Is this enough?

Re-identification by Linking

Microdata

ID	QID			SA
Name	Zipcode	Age	Sex	Disease
Alice	47677	29	F	Ovarian Cancer
Betty	47602	22	F	Ovarian Cancer
Charles	47678	27	M	Prostate Cancer
David	47905	43	M	Flu
Emily	47909	52	F	Heart Disease
Fred	47906	47	M	Heart Disease

Voter registration data

Name	Zipcode	Age	Sex
Alice	47677	29	F
Bob	47983	65	M
Carol	47677	22	F
Dan	47532	23	M
Ellen	46789	43	F

Latanya Sweeney's Attack (1997)

Massachusetts hospital discharge dataset

Medical Data Released as Anonymous

SSN	Name	Ethnicity	Date Of Birth	Sex	ZIP	Marital Status	Problem
			09/27/64	female	02139	divorced	hypertension
			09/30/64	female	02139	divorced	obesity
		asian	04/18/64	male	02139	married	chest pain
		asian	04/15/64	male	02139	married	obesity
		black	03/13/63	male	02138	married	hypertension
		black	03/18/63	male	02138	married	shortness of breath
		black	09/13/64	female	02141	married	shortness of breath
		black	09/07/64	female	02141	married	obesity
		white	05/14/61	male	02138	single	chest pain
		white	05/08/61	male	02138	single	obesity
		white	09/15/61	female	02142	widow	shortness of breath

Voter List

Name	Address	City	ZIP	DOB	Sex	Party	....
....	....	....	....	....	....	....	
....	....	....	....	....	....	....	
Sue J. Carlson	1459 Main St.	Cambridge	*02142*	*9/15/61*	*female*	democrat	....
....	....	....	....	....	....	....	

Figure 1: Re-identifying anonymous data by linking to external data

Public voter dataset

Quasi-Identifiers

- Key attributes
  - Name, address, phone number - uniquely identifying!
  - Always removed before release
- Quasi-identifiers
  - (5-digit ZIP code, birth date, gender) uniquely identify 87% of the population in the U.S.
  - Can be used for linking anonymized dataset with other datasets

Classification of Attributes

- **Sensitive attributes**
  - Medical records, salaries, etc.
  - These attributes is what the researchers need, so they are always released directly

Key Attribute	Quasi-identifier			Sensitive attribute
Name	DOB	Gender	Zipcode	Disease
Andre	1/21/76	Male	53715	Heart Disease
Beth	4/13/86	Female	53715	Hepatitis
Carol	2/28/76	Male	53703	Brochitis
Dan	1/21/76	Male	53703	Broken Arm
Ellen	4/13/86	Female	53706	Flu
Eric	2/28/76	Female	53706	Hang Nail

K-Anonymity: Intuition

- The information for each person contained in the released table cannot be distinguished from at least k-1 individuals whose information also appears in the release
  - Example: you try to identify a man in the released table, but the only information you have is his birth date and gender. There are k men in the table with the same birth date and gender.
- Any quasi-identifier present in the released table must appear in at least k records

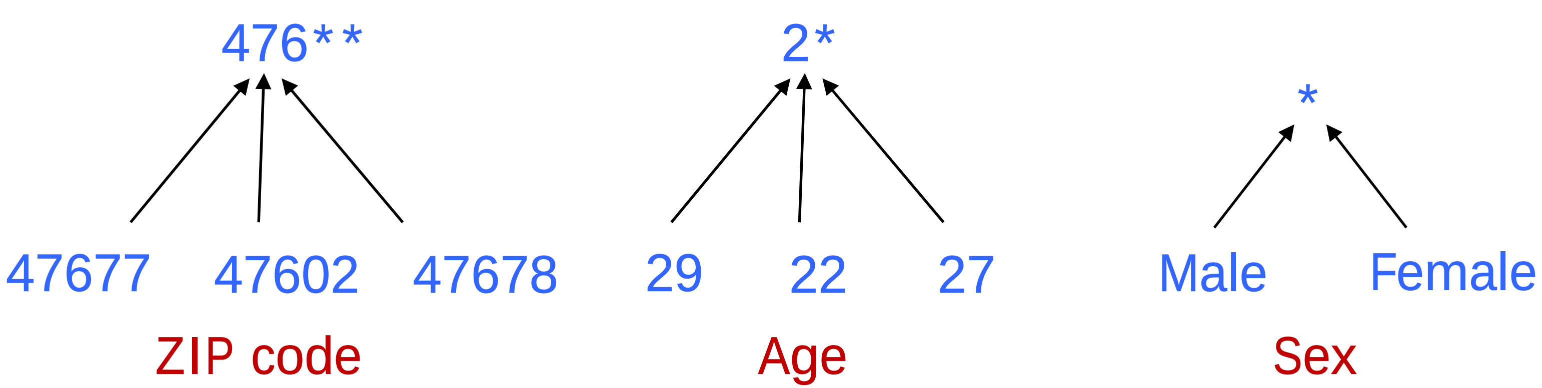
K-Anonymity Protection Model

- Private table
- Released table: RT
- Attributes: $A_1, A_2, \ldots, A_n$
- Quasi-identifier subset: $A_i, \ldots, A_j$

Let RT($A_1,\ldots,A_n$) be a table, $QI_{RT} = (A_i, \ldots, A_j)$ be the quasi-identifier associated with RT, $A_i, \ldots A_j \subseteq A_1, \ldots, A_n$, and RT satisfy k-anonymity. Then, each sequence of values in RT[A_x] appears with at least k occurrences in RT[QI_{RT}] for $x=i,\ldots,j$.

Generalization

- Goal of k-Anonymity
  - Each record is indistinguishable from at least k-1 other records
  - These k records form an equivalence class
- Generalization: replace quasi-identifiers with less specific, but semantically consistent values

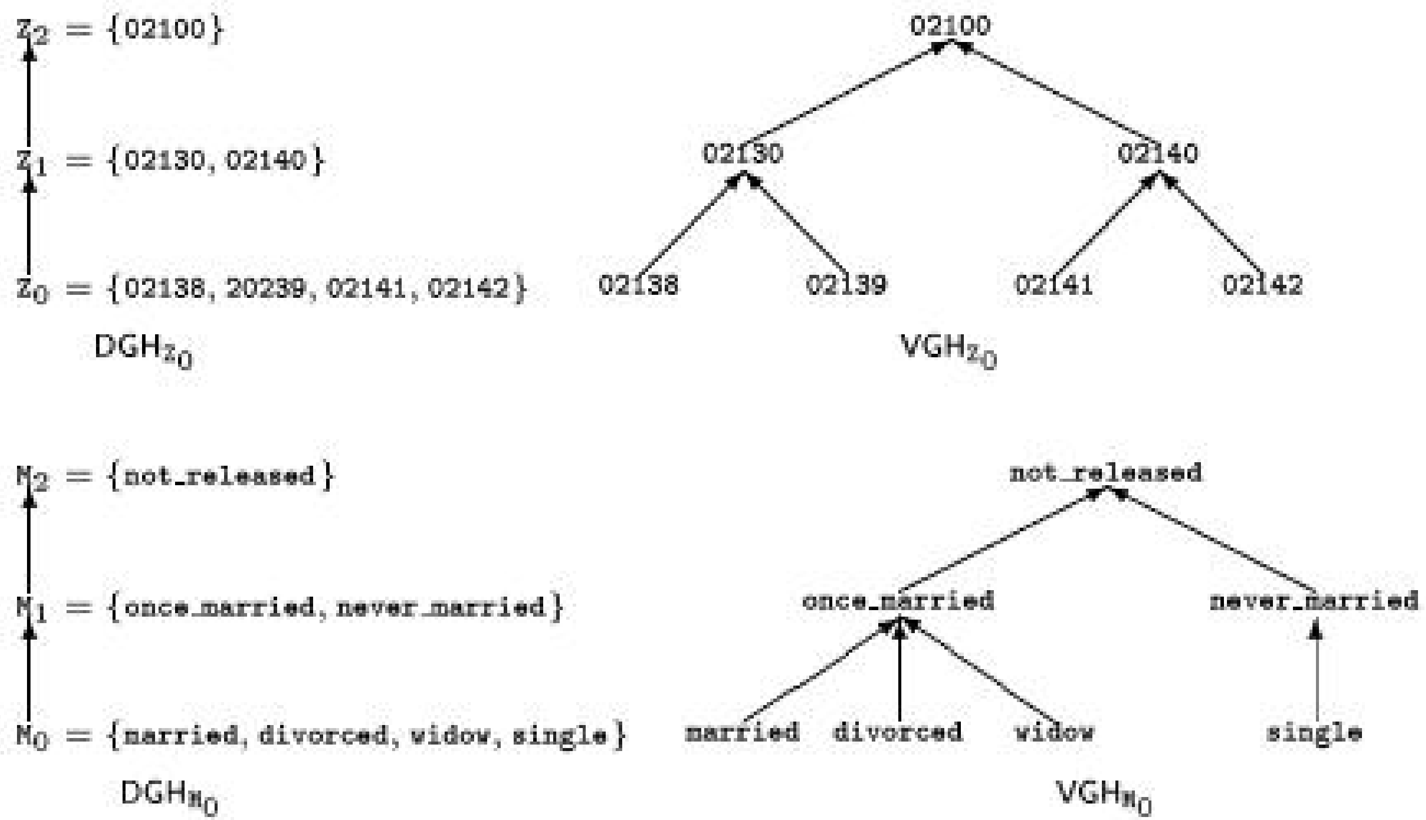


Achieving k-Anonymity

- Generalization
  - Replace specific quasi-identifiers with less specific values until get k identical values
  - Partition ordered-value domains into intervals
- Suppression
  - When generalization causes too much information loss
    - This is common with “outliers”
- Lots of algorithms in the literature
  - Aim to produce “useful” anonymizations

    ... usually without any clear notion of utility

Generalization in Action



Example of a k-Anonymous Table

	Race	Birth	Gender	ZIP	Problem
t1	Black	1965	m	0214*	short breath
t2	Black	1965	m	0214*	chest pain
t3	Black	1965	f	0213*	hypertension
t4	Black	1965	f	0213*	hypertension
t5	Black	1964	f	0213*	obesity
t6	Black	1964	f	0213*	chest pain
t7	White	1964	m	0213*	chest pain
t8	White	1964	m	0213*	obesity
t9	White	1964	m	0213*	short breath
t10	White	1967	m	0213*	chest pain
t11	White	1967	m	0213*	chest pain

Figure 2 Example of *k*-anonymity, where *k*=2 and QI={*Race*, *Birth*, *Gender*, *ZIP*}

Example of Generalization (1)

Released table

	Race	Birth	Gender	ZIP	Problem
t1	Black	1965	m	0214*	short breath
t2	Black	1965	m	0214*	chest pain
t3	Black	1965	f	0213*	hypertension
t4	Black	1965	f	0213*	hypertension
t5	Black	1964	f	0213*	obesity
t6	Black	1964	f	0213*	chest pain
t7	White	1964	m	0213*	chest pain
t8	White	1964	m	0213*	obesity
t9	White	1964	m	0213*	short breath
t10	White	1967	m	0213*	chest pain
t11	White	1967	m	0213*	chest pain

External data

Name	Birth	Gender	ZIP	Race
Andre	1964	m	02135	White
Beth	1964	f	55410	Black
Carol	1964	f	90210	White
Dan	1967	m	02174	White
Ellen	1968	f	02237	White

By linking these 2 tables, you still don't learn Andre's problem

Example of Generalization (2)

Microdata

QID			SA
Zipcode	Age	Sex	Disease
47677	29	F	Ovarian Cancer
47602	22	F	Ovarian Cancer
47678	27	M	Prostate Cancer
47905	43	M	Flu
47909	52	F	Heart Disease
47906	47	M	Heart Disease

Generalized table

QID			SA
Zipcode	Age	Sex	Disease
476**	2*	*	Ovarian Cancer
476**	2*	*	Ovarian Cancer
476**	2*	*	Prostate Cancer
4790*	[43,52]	*	Flu
4790*	[43,52]	*	Heart Disease
4790*	[43,52]	*	Heart Disease

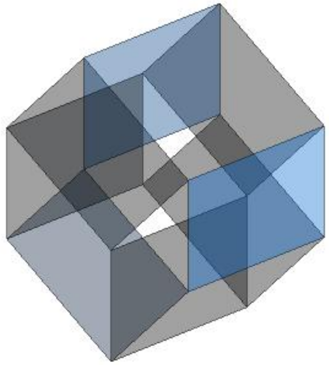
!!

- Released table is 3-anonymous
- If the adversary knows Alice's quasi-identifier (47677, 29, F), he still does not know which of the first 3 records corresponds to Alice's record

Curse of Dimensionality

[Aggarwal VLDB '05]

- Generalization fundamentally relies on spatial locality
  - Each record must have k close neighbors
- Real-world datasets are very sparse
  - Many attributes (dimensions)
    - Netflix Prize dataset: 17,000 dimensions
    - Amazon customer records: several million dimensions
  - "Nearest neighbor" is very far
- Projection to low dimensions loses all info $\Rightarrow$ k-anonymized datasets are useless



HIPAA Privacy Rule

"Under the safe harbor method, covered entities must remove all of a list of 18 enumerated identifiers and have no actual knowledge that the information remaining could be used, alone or in combination, to identify a subject of the information."

“The identifiers that must be removed include direct identifiers, such as name, street address, social security number, as well as other identifiers, such as birth date, admission and discharge dates, and five-digit zip code. The safe harbor requires removal of geographic subdivisions smaller than a State, except for the initial three digits of a zip code if the geographic unit formed by combining all zip codes with the same initial three digits contains more than 20,000 people. In addition, age, if less than 90, gender, ethnicity, and other demographic information not listed may remain in the information. The safe harbor is intended to provide covered entities with a simple, definitive method that does not require much judgment by the covered entity to determine if the information is adequately de-identified."

Two (and a Half) Interpretations

- **Membership disclosure:** Attacker cannot tell that a given person in the dataset
- **Sensitive attribute disclosure:** Attacker cannot tell that a given person has a certain sensitive attribute
- **Identity disclosure:** Attacker cannot tell which record corresponds to a given person

This interpretation is correct, assuming the attacker does not know anything other than quasi-identifiers

But this does not imply any privacy!

Example: k clinical records, all HIV+

Unsorted Matching Attack

- Problem: records appear in the same order in the released table as in the original table
- Solution: randomize order before releasing

Race	ZIP
Asian	02138
Asian	02139
Asian	02141
Asian	02142
Black	02138
Black	02139
Black	02141
Black	02142
White	02138
White	02139
White	02141
White	02142

PT

Race	ZIP
Person	02138
Person	02139
Person	02141
Person	02142
Person	02138
Person	02139
Person	02141
Person	02142
Person	02138
Person	02139
Person	02141
Person	02142

GT1

Race	ZIP
Asian	02130
Asian	02130
Asian	02140
Asian	02140
Black	02130
Black	02130
Black	02140
Black	02140
White	02130
White	02130
White	02140
White	02140

GT2

Figure 3 Examples of *k*-anonymity tables based on PT

Complementary Release Attack

- Different releases of the same private table can be linked together to compromise k-anonymity

Race	BirthDate	Gender	ZIP	Problem
black	1965	male	02141	short of breath
black	1965	male	02141	chest pain
person	1965	female	0213*	painful eye
person	1965	female	0213*	wheezing
black	1964	female	02138	obesity
black	1964	female	02138	chest pain
white	1964	male	0213*	short of breath
person	1965	female	0213*	hypertension
white	1964	male	0213*	obesity
white	1964	male	0213*	fever
white	1967	male	02138	vomiting
white	1967	male	02138	back pain

GT1

Race	BirthDate	Gender	ZIP	Problem
black	1965	male	02141	short of breath
black	1965	male	02141	chest pain
black	1965	female	02138	painful eye
black	1965	female	02138	wheezing
black	1964	female	02138	obesity
black	1964	female	02138	chest pain
white	1960-69	male	02138	short of breath
white	1960-69	human	02139	hypertension
white	1960-69	human	02139	obesity
white	1960-69	human	02139	fever
white	1960-69	male	02138	vomiting
white	1960-69	male	02138	back pain

GT3

Linking Independent Releases

Race	BirthDate	Gender	ZIP	Problem
black	9/20/1965	male	02141	short of breath
black	2/14/1965	male	02141	chest pain
black	10/23/1965	female	02138	painful eye
black	8/24/1965	female	02138	wheezing
black	11/7/1964	female	02138	obesity
black	12/1/1964	female	02138	chest pain
white	10/23/1964	male	02138	short of breath
white	3/15/1965	female	02139	hypertension
white	8/13/1964	male	02139	obesity
white	5/5/1964	male	02139	fever
white	2/13/1967	male	02138	vomiting
white	3/21/1967	male	02138	back pain

PT

Race	BirthDate	Gender	ZIP	Problem
black	1965	male	02141	short of breath
black	1965	male	02141	chest pain
black	1965	female	02138	painful eye
black	1965	female	02138	wheezing
black	1964	female	02138	obesity
black	1964	female	02138	chest pain
white	1964	male	02138	short of breath
white	1965	female	02139	hypertension
white	1964	male	02139	obesity
white	1964	male	02139	fever
white	1967	male	02138	vomiting
white	1967	male	02138	back pain

LT

Attacks on k-Anonymity

- k-Anonymity does not provide privacy if
  - Sensitive values in an equivalence class lack diversity
  - The attacker has background knowledge

Homogeneity attack

Bob	
Zipcode	***Age***
47678	27

Background knowledge attack

Carl	
Zipcode	***Age***
47673	36

A 3-anonymous patient table

Zipcode	Age	Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
476**	2*	Heart Disease
4790*	≥40	Flu
4790*	≥40	Heart Disease
4790*	≥40	Cancer
476**	3*	Heart Disease
476**	3*	Cancer
476**	3*	Cancer

l-Diversity

[Machanavajjhala et al. ICDE '06]

Caucas	787XX	Flu
Caucas	787XX	Shingles
Caucas	787XX	Acne
Caucas	787XX	Flu
Caucas	787XX	Acne
Caucas	787XX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Shingles
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Flu

Sensitive attributes must be "diverse" within each quasi-identifier equivalence class

Distinct l-Diversity

- Each equivalence class has at least l well-represented sensitive values
- Doesn't prevent probabilistic inference attacks

…	**Disease**
	…
	HIV
	HIV
	…
	HIV
	pneumonia
	bronchitis
	…

10 records

8 records have HIV

2 records have other values

Other Versions of l-Diversity

- Probabilistic l-diversity
  - The frequency of the most frequent value in an equivalence class is bounded by 1/l
- Entropy l-diversity
  - The entropy of the distribution of sensitive values in each equivalence class is at least log(l)
- Recursive (c,l)-diversity
  - $r_1 < c(r_l + r_{l+1} + \ldots + r_m)$ where r_i is the frequency of the i^{th} most frequent value
  - Intuition: the most frequent value does not appear too frequently

Neither Necessary, Nor Sufficient

Original dataset

…	Cancer
…	Cancer
…	Cancer
…	Flu
…	Cancer
…	Cancer
…	Cancer
…	Cancer
…	Cancer
…	Cancer
…	Flu
…	Flu

99% have cancer

Anonymization A

Q1	Flu
Q1	Flu
Q1	Cancer
Q1	Flu
Q1	Cancer
Q1	Cancer
Q2	Cancer
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
Q	[illegible]

Anonymization B

Q1	Flu
Q1	Cancer
Q1	Cancer
Q1	Cancer
Q1	Cancer
Q1	Cancer
Q2	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
[illegible]	[illegible]
Q2	Flu

99% cancer ⇒ quasi-identifier group is <u>not</u> "diverse"
…yet anonymized database does not leak anything

50% cancer ⇒ quasi-identifier group is "diverse"
This leaks a ton of information

Limitations of l-Diversity

- Example: sensitive attribute is HIV+ (1%) or HIV- (99%)
  - Very different degrees of sensitivity!
- l-diversity is unnecessary
  - 2-diversity is unnecessary for an equivalence class that contains only HIV- records
- l-diversity is difficult to achieve
  - Suppose there are 10000 records in total
  - To have distinct 2-diversity, there can be at most 10000*1%=100 equivalence classes

Skewness Attack

- Example: sensitive attribute is HIV+ (1%) or HIV- (99%)
- Consider an equivalence class that contains an equal number of HIV+ and HIV- records
  - Diverse, but potentially violates privacy!
- l-diversity does not differentiate:
  - Equivalence class 1: 49 HIV+ and 1 HIV-
  - Equivalence class 2: 1 HIV+ and 49 HIV-

l-diversity does not consider overall distribution of sensitive values!

Sensitive Attribute Disclosure

Similarity attack

Bob	
Zip	***Age***
47678	27

A 3-diverse patient table

Zipcode	Age	Salary	Disease
476**	2*	20K	Gastric Ulcer
476**	2*	30K	Gastritis
476**	2*	40K	Stomach Cancer
4790*	≥40	50K	Gastritis
4790*	≥40	100K	Flu
4790*	≥40	70K	Bronchitis
476**	3*	60K	Bronchitis
476**	3*	80K	Pneumonia
476**	3*	90K	Stomach Cancer

Conclusion

1. Bob's salary is in [20k,40k], which is relatively low
2. Bob has some stomach-related disease

l-diversity does not consider semantics of sensitive values!

t-Closeness

[Li et al. ICDE '07]

Caucas	787XX	Flu
Caucas	787XX	Shingles
Caucas	787XX	Acne
Caucas	787XX	Flu
Caucas	787XX	Acne
Caucas	787XX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Flu
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Shingles
Asian/AfrAm	78XXX	Acne
Asian/AfrAm	78XXX	Flu

Distribution of sensitive attributes within each quasi-identifier group should be "close" to their distribution in the entire original database

Trick question: Why publish quasi-identifiers at all??

Anonymous, “t-Close” Dataset

Caucas	787XX	HIV+	Flu
Asian/AfrAm	787XX	HIV-	Flu
Asian/AfrAm	787XX	HIV+	Shingles
Caucas	787XX	HIV-	Acne
Caucas	787XX	HIV-	Shingles
Caucas	787XX	HIV-	Acne

This is k-anonymous, l-diverse and t-close...

...so secure, right?

What Does Attacker Know?

Bob is Caucasian and
I heard he was
admitted to hospital
with flu…

This is against the rules!
"flu" is not a quasi-identifier

Yes… and this is yet another
problem with k-anonymity

Caucas	787XX	HIV+	Flu
Asian/AfrAm	787XX	HIV-	Flu
[illegible]	[illegible]	HIV+	Shingles
Caucas	787XX	HIV-	Acne
[illegible]	[illegible]	HIV-	Shingles
Caucas	787XX	HIV-	Acne

AOL Privacy Debacle

- In August 2006, AOL released anonymized search query logs
  - 657K users, 20M queries over 3 months (March-May)
- Opposing goals
  - Analyze data for research purposes, provide better services for users and advertisers
  - Protect privacy of AOL users
    - Government laws and regulations
    - Search queries may reveal income, evaluations, intentions to acquire goods and services, etc.

AOL User 4417749

- AOL query logs have the form

  <AnonID, Query, QueryTime, ItemRank, ClickURL>

  - ClickURL is the truncated URL
- NY Times re-identified AnonID 4417749
  - Sample queries: "numb fingers", "60 single men", "dog that urinates on everything", "landscapers in Lilburn, GA", several people with the last name Arnold
    - Lilburn area has only 14 citizens with the last name Arnold
  - NYT contacts the 14 citizens, finds out AOL User 4417749 is 62-year-old Thelma Arnold

k-Anonymity Considered Harmful

- Syntactic
  - Focuses on data transformation, not on what can be learned from the anonymized dataset
  - “k-anonymous” dataset can leak sensitive information
- “Quasi-identifier” fallacy
  - Assumes a priori that attacker will not know certain information about his target
- Relies on locality
  - Destroys utility of many real-world datasets