Progress in Data Anonymization: from k-anonymity to the minimality attack

Graham Cormode

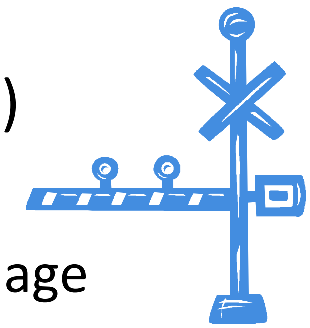
graham@research.att.com

Tiancheng Li, Ninghua Li, Divesh Srivastava

Why Anonymize?

- For Data Sharing
  - Give real(istic) data to others to study without compromising privacy of individuals in the data
  - Allows third-parties to try new analysis and mining techniques not thought of by the data owner
- For Data Retention and Usage
  - Various requirements prevent companies from retaining customer information indefinitely
  - E.g. Google progressively anonymizes IP addresses in search logs
  - Internal sharing across departments (e.g. billing $\rightarrow$ marketing)

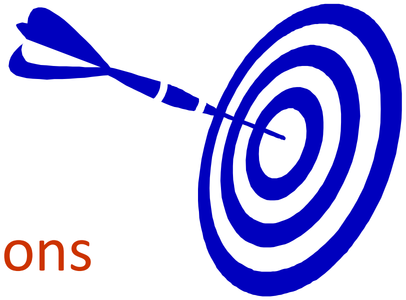
Models of Anonymization



- Interactive Model (akin to statistical databases)
  - Data owner acts as “gatekeeper” to data
  - Researchers pose queries in some agreed language
  - Gatekeeper gives an (anonymized) answer, or refuses to answer
- “Send me your code” model
  - Data owner executes code on their system and reports result
  - Cannot be sure that the code is not malicious, compiles…
- Offline, aka “publish and be damned” model
  - Data owner somehow anonymizes data set
  - Publishes the results, and retires
  - Seems to best model many real releases

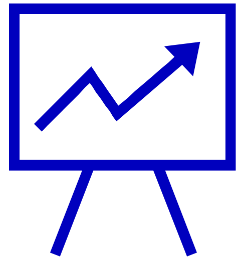
at&t

Objectives for Anonymization



- Prevent (high confidence) inference of associations
  - Prevent inference of salary for an individual in census data
  - Prevent inference of individual’s video viewing history
  - Prevent inference of individual’s search history in search logs
  - All aim to prevent linking sensitive information to an individual
- Have to model what knowledge might be known to attacker
  - Background knowledge: facts about the data set (X has salary Y)
  - Domain knowledge: broad properties of data (illness Z rare in men)

Utility



- Anonymization is meaningless if utility of data not considered
  - The empty data set has perfect privacy, but no utility
  - The original data has full utility, but no privacy
- What is "utility"? Depends what the application is…
  - For fixed query set, can look at max, average distortion
  - Problem for publishing: want to support unknown applications!
  - Need some way to quantify utility of alternate anonymizations

Outline

- Introduction to Anonymization
- Linking Attack and k-anonymization
- Homogeneity Attack and l-diversity
- Minimality Attack and analysis

Tabular Data Example

- Census data recording incomes and demographics

SSN	DOB	Sex	ZIP	Salary
11-1-111	1/21/76	M	53715	50,000
22-2-222	4/13/86	F	53715	55,000
33-3-333	2/28/76	M	53703	60,000
44-4-444	1/21/76	M	53703	65,000
55-5-555	4/13/86	F	53706	70,000
66-6-666	2/28/76	F	53706	75,000

- Releasing SSN $\rightarrow$ Salary association violates individual's privacy
  - SSN is an identifier, Salary is a sensitive attribute (SA)

Tabular Data Example: De-Identification

- Census data: remove SSN to create de-identified table

DOB	Sex	ZIP	Salary
1/21/76	M	53715	50,000
4/13/86	F	53715	55,000
2/28/76	M	53703	60,000
1/21/76	M	53703	65,000
4/13/86	F	53706	70,000
2/28/76	F	53706	75,000

- Does the de-identified table preserve an individual's privacy?
  - Depends on what other information an attacker knows

Tabular Data Example: Linking Attack

- De-identified private data + publicly available data

DOB	Sex	ZIP	Salary
1/21/76	M	53715	50,000
4/13/86	F	53715	55,000
2/28/76	M	53703	60,000
1/21/76	M	53703	65,000
4/13/86	F	53706	70,000
2/28/76	F	53706	75,000

SSN	DOB
11-1-111	1/21/76
33-3-333	2/28/76

- Cannot uniquely identify either individual's salary
  - DOB is a quasi-identifier (QI)

Tabular Data Example: Linking Attack

- De-identified private data + publicly available data

DOB	Sex	ZIP	Salary
1/21/76	M	53715	50,000
4/13/86	F	53715	55,000
2/28/76	M	53703	60,000
1/21/76	M	53703	65,000
4/13/86	F	53706	70,000
2/28/76	F	53706	75,000

SSN	DOB	Sex
11-1-111	1/21/76	M
33-3-333	2/28/76	M

- Uniquely identified one individual's salary, but not the other's
  - DOB, Sex are quasi-identifiers (QI)

Tabular Data Example: Linking Attack

- De-identified private data + publicly available data

DOB	Sex	ZIP	Salary
1/21/76	M	53715	50,000
4/13/86	F	53715	55,000
2/28/76	M	53703	60,000
1/21/76	M	53703	65,000
4/13/86	F	53706	70,000
2/28/76	F	53706	75,000

SSN	DOB	Sex	ZIP
11-1-111	1/21/76	M	53715
33-3-333	2/28/76	M	53703

- Uniquely identified both individuals' salaries
  - [DOB, Sex, ZIP] is unique for majority of US residents [Sweeney 02]

Tabular Data Example: Anonymization

- Anonymization through QI attribute generalization

DOB	Sex	ZIP	Salary
1/21/76	M	537**	50,000
4/13/86	F	537**	55,000
2/28/76	*	537**	60,000
1/21/76	M	537**	65,000
4/13/86	F	537**	70,000
2/28/76	*	537**	75,000

SSN	DOB	Sex	ZIP
11-1-111	1/21/76	M	53715
33-3-333	2/28/76	M	53703

- Cannot uniquely identify tuple with knowledge of QI values
  - E.g., ZIP = 537** → ZIP ∈ {53700, …, 53799}

Tabular Data Example: Anonymization

- Anonymization through sensitive attribute (SA) permutation

DOB	Sex	ZIP	Salary
1/21/76	M	53715	55,000
4/13/86	F	53715	50,000
2/28/76	M	53703	60,000
1/21/76	M	53703	65,000
4/13/86	F	53706	75,000
2/28/76	F	53706	70,000

SSN	DOB	Sex	ZIP
11-1-111	1/21/76	M	53715
33-3-333	2/28/76	M	53703

- Can uniquely identify tuple, but uncertainty about SA value
  - Much more precise form of uncertainty than generalization

k-Anonymization [Samarati, Sweeney 98]

- k-anonymity: Table T satisfies k-anonymity wrt quasi-identifiers QI iff each tuple in (the multiset) T[QI] appears at least k times
  - Protects against “linking attack”
- k-anonymization: Table T’ is a k-anonymization of T if T’ is generated from T, and T’ satisfies k-anonymity

DOB	Sex	ZIP	Salary
1/21/76	M	53715	50,000
4/13/86	F	53715	55,000
2/28/76	M	53703	60,000
1/21/76	M	53703	65,000
4/13/86	F	53706	70,000
2/28/76	F	53706	75,000

→

DOB	Sex	ZIP	Salary
1/21/76	M	537**	50,000
4/13/86	F	537**	55,000
2/28/76	*	537**	60,000
1/21/76	M	537**	65,000
4/13/86	F	537**	70,000
2/28/76	*	537**	75,000

Outline

- Introduction to Anonymization
- Linking Attack and k-anonymization
- Homogeneity Attack and l-diversity
- Minimality Attack and analysis

Homogeneity Attack [Machanavajjhala+ 06]

- Issue: k-anonymity requires each tuple in (the multiset) T[QI] to appear ≥ k times, but does not say anything about the SA values
  - If (almost) all SA values in a QI group are equal, loss of privacy!
  - The problem is with the choice of grouping, not the data
  - For some groupings, no loss of privacy

DOB	Sex	ZIP	Salary
1/21/76	M	53715	50,000
4/13/86	F	53715	55,000
2/28/76	M	53703	60,000
1/21/76	M	53703	50,000
4/13/86	F	53706	55,000
2/28/76	F	53706	60,000

Ok! →

DOB	Sex	ZIP	Salary
76-86	*	53715	50,000
76-86	*	53715	55,000
76-86	*	53703	60,000
76-86	*	53703	50,000
76-86	*	53706	55,000
76-86	*	53706	60,000

l-Diversity [Machanavajjhala+ 06]

- Intuition: Most frequent value does not appear too often compared to the less frequent values in a QI group
- Simplified *l*-diversity defn: for each group, max frequency $\leq 1/l$
  - *l*-diversity((1/21/76, *, 537**)) = 1

DOB	Sex	ZIP	Salary
1/21/76	*	537**	50,000
4/13/86	*	537**	55,000
2/28/76	*	537**	60,000
1/21/76	*	537**	50,000
4/13/86	*	537**	55,000
2/28/76	*	537**	60,000

Simple Algorithm for *l*-diversity

- Simple "Greedy Grouping" algorithm provides *l*-diversity
  - Sort tuples based on attributes so similar tuples are close
  - Start with group containing just first tuple
  - Keeping adding tuples to group in order until l-diversity met
  - Output the group, and repeat on remaining tuples

DOB	Sex	ZIP	Salary
1/21/76	M	53715	50,000
4/13/86	F	53715	50,000
2/28/76	M	53703	60,000
1/21/76	M	53703	65,000
4/13/86	F	53706	50,000
2/28/76	F	53706	60,000

2-diversity →

DOB	Sex	ZIP	Salary
1/21/76	M	53715	50,000
4/13/86	F	53715	50,000
2/28/76	M	53703	60,000
1/21/76	M	53703	65,000
4/13/86	F	53706	50,000
2/28/76	F	53706	60,000

Outline

- Introduction to Anonymization
- Linking Attack and k-anonymization
- Homogeneity Attack and l-diversity
- Minimality Attack and analysis

Minimality Attack [Wong Fu Wang Pei 07]

- In *l*-diversity analysis, we assume that all possible inputs consistent with the output are equally likely
- Minimality attack: condition on knowledge of the algorithm
  - Some inputs would not have resulted in that output

DOB	Sex	ZIP	Salary
1/21/76	M	53715	50,000
4/13/86	F	53715	60,000
2/28/76	M	53703	50,000
1/21/76	M	53703	65,000
4/13/86	F	53706	50,000
2/28/76	F	53706	60,000

2-diversity →

DOB	Sex	ZIP	Salary
1/21/76	M	53715	50,000
4/13/86	F	53715	60,000
2/28/76	M	53703	50,000
1/21/76	M	53703	65,000
4/13/86	F	53706	50,000
2/28/76	F	53706	60,000

Minimality attack

- In our example, can use knowledge of anonymization algorithm to learn sensitive values!
  - No additional knowledge needed
- More generally, attacker associates a probability with each tuple and each sensitive value
  - l-diversity: this probability should be at most 1/l
  - Using minimality attack, this probability exceeds 1/l
- Our goal: understand this attack better
  - Can the attack inflate probabilities arbitrarily?

Binary l-diversity

- For simplicity, study a special case: sensitive attribute is binary
  - (binary) l-diversity: each group should have at most 1/l fraction of positive values
  - Safe to have a group of all negative values

DOB	Sex	ZIP	Disease
1/21/76	M	53715	N
4/13/86	F	53715	Y
2/28/76	M	53703	N
1/21/76	M	53703	N
4/13/86	F	53706	Y
2/28/76	F	53706	N

Analysis of Greedy Grouping + Minimality

- Consider each group output by GG independently
  - Fraction of positive tuples ≤ 1/l (by l-diversity)
  - By minimality, each prefix of group has > 1/l positive tuples
  - First l tuples must have ≥ 2 positives → prob on these is ≥ 2/l
- Divide group into each "bucket" of l tuples
  - Cannot distinguish between tuples in each bucket
  - Each bucket b has an associated probability, p(b)

DOB	Sex	ZIP	Disease	
1/21/76	M	53715	N	b_1
4/13/86	F	53715	Y	
2/28/76	M	53703	N	
1/21/76	M	53703	N	b_2
4/13/86	F	53706	Y	
2/28/76	F	53706	N	

l=3

Reduction to First Bucket

- Claim: first bucket b_1 has highest probability in group
  - Consider all possible worlds that have n_i positives in bucket i
  - Over m buckets, must have $n_1 + n_2 + \ldots + n_m = m$ (by l-diversity)
  - Consider buckets b_i, and b_{i+1}, holding other n_i's constant
  - Then $n_i + n_{i+1} = r$ must be fixed
  - Let t denote smallest value of n_i that gives a valid sequence
- Let N_j denote number of worlds with $(n_i = j, n_{i+1} = r-j)$
  - Then $N_j = N_{r-j}$ – can give bijection by swapping buckets
- Calculate probabilities for each bucket
  - $p(b_i) = \sum_{j=t}^{r} j\, N_j / \sum_{j=t}^{r} N_j$ and $p(b_{i+1}) = \sum_{i=t}^{r} (r-j)\, N_j / \sum_{j=t}^{r} N_j$

Analysis of bucket probabilities

- First bucket is most probable if $p(b_i) - p(b_{i+1}) \geq 0$ for all i
  - We have $p(b_i) - p(b_{i+1}) = \sum_{i=t}^{r} (2j - r)N_j / \sum_{j=t} N_j$
- When $t > r/2$, $(2j - r)N_j$ for all j
- For $t < r/2$, split the numerator into pieces around $r/2$:
  - $p(b_i) - p(b_{i+1}) = T + \sum_{j=t}^{\lfloor r/2 \rfloor} (2j\text{-}r)N_j + \sum_{j=\lfloor r/2 \rfloor + 1}^{r\text{-}t} (2j - r) N_j$
  - $= T + \sum_{j=r\text{-}\lfloor r/2 \rfloor}^{r\text{-}t} (2(r\text{-}j)\text{-}r)N_{r\text{-}j} + \sum_{j=\lfloor r/2 \rfloor + 1}^{r\text{-}t} (2j - r) N_j$ [swap N_j for $N_{r\text{-}j}$]
  - $= T + \sum_{j=\lfloor r/2 \rfloor + 1}^{r\text{-}t} (r\text{-}2j)N_j + \sum_{j=\lfloor r/2 \rfloor + 1}^{r\text{-}t} (2j - r) N_j$ [rearrange]
  - $= T \geq 0$
- This proves the claim that highest probability is in first bucket

Analyze first bucket probability

- Let $m^* = \min\{m, l\}$, an upper bound on any n_i
- Let N_k be number of possible worlds where $n_1 = k$
- Then $p(b_1) = \sum_{k=2}^{m^*} k\ N_k \,/\, \sum_{k=2}^{m^*} l\ N_k$
  - Expected fraction of positives in first bucket
- Sequence of steps to analyze $p(b_1)$:
  - Compute N_k
  - Compute the numerator $\sum_{k=2}^{m^*} k\ N_k$
  - Compute the denominator $\sum_{k=2}^{m^*} l\ N_k$

Analysis of N_k

- Consider all sequences of n_i's which begin with k
  - $(k, n_2, n_3, \ldots, n_m)$
  - Consider all permutations of this set of n_i's which keep k first
  - Validity: must have $k + \sum_{i=2}^{j} n_i > j$ for $j<m$ else prefix is l-diverse
- Claim: exactly a $(k-1)/(m-1)$ fraction of permutations are valid
  - Proof by induction on m
  - **Base case**: any permutation of $(m, 0, \ldots 0)$ is valid
  - **Inductive case**: build a $m+1$ valid sequence from m sequences
  - Connection to Catalan numbers and Dyck paths

Computing N_k

- Consider how to make a world with $n_1 = k$
  - Pick k positions from first bucket to be positive
  - Place rest of positive items in rest of positions
  - Gives $(l \text{ C } k)\ ((ml - l) \text{ C } (m-k))$ possibilities
- Exactly a $(k-1)/(m-1)$ fraction of these are 'valid'
  - So $N_k = (k-1)/(m-1)\ (l \text{ C } k)\ ((ml - l) \text{ C } (m-k))$

Computing Numerator and Denominator

- Numerator:

$$\sum_{k=2}^{m^*} k\, N_k = \sum_{k=2}^{m^*} k(k-1)/(m-1)\ (l\ C\ k)(ml - l\ C\ m-k)$$
$$= \sum_{k=2}^{m^*} l(l-1)/(m-1)\ (l-2\ C\ k-2)(ml - l\ C\ m-k)$$
$$= l\,(l-1)/(m-1)\ (ml -2\ C\ m-2)$$

- Denominator:

$$(m-1)\sum_{k=2}^{m^*} N_k = \sum_{k=2}^{m^*} (k-1)/(m-1)\ (l\ C\ k)(ml -1\ C\ m-k)$$
$$= \sum_{k=1}^{m^*} k(l\ C\ k)(ml -1\ C\ m-k) - (l\ C\ k)(ml - 1\ C\ m-k)$$
$$= \sum_{k=1}^{m^*} l\ (l-1\ C\ k-1)\ (ml - l\ C\ m-k) - ((ml\ C\ m) - (ml - l\ C\ M))$$
$$= l\ (ml -1\ C\ m-1) - (ml\ C\ m) + (ml - 1\ C\ m)$$
$$= (ml - l\ C\ m)$$

Bounding $p(b_l)$

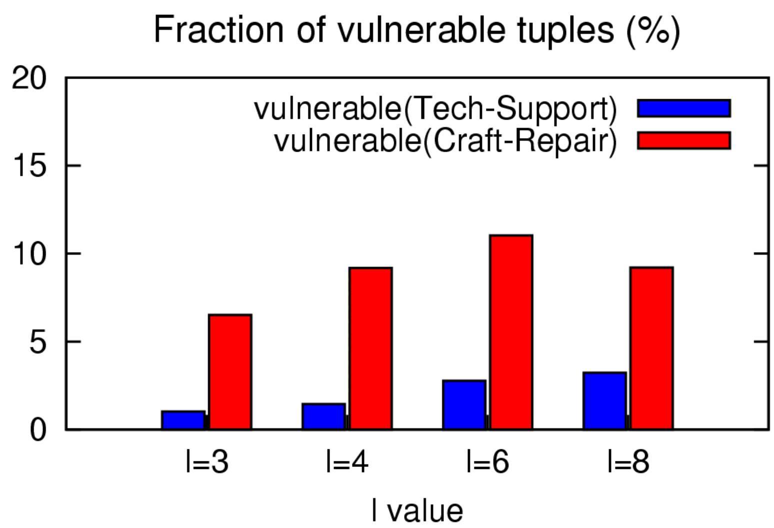
- $p(b_1)$ $= (l\text{-}1)(ml\ \text{-}2\ C\ m\text{-}2)/(ml - l\ C\ m)$
  $= (ml\ \text{-}2)!\ (ml - l - m)!/(\ l(ml\text{-}m)!\ (ml\text{-}l)!\ )$
  $= \prod_{j=1}^{l-1} (ml\text{-}1\text{-}j)/l(ml\text{-}m\text{-}j)$
  $= \prod_{j=1}^{l-1} (1 + (m\text{-}1)/(ml\text{-}m\text{-}j))/l$
  $\leq \prod_{j=1}^{l-1} (1 + (m\text{-}1)/(ml\text{-}m\text{-}(l\text{-}1)))/l$
  $= 1/l\ (1 + 1/(l\text{-}1))^{(l-1)} < e/l$

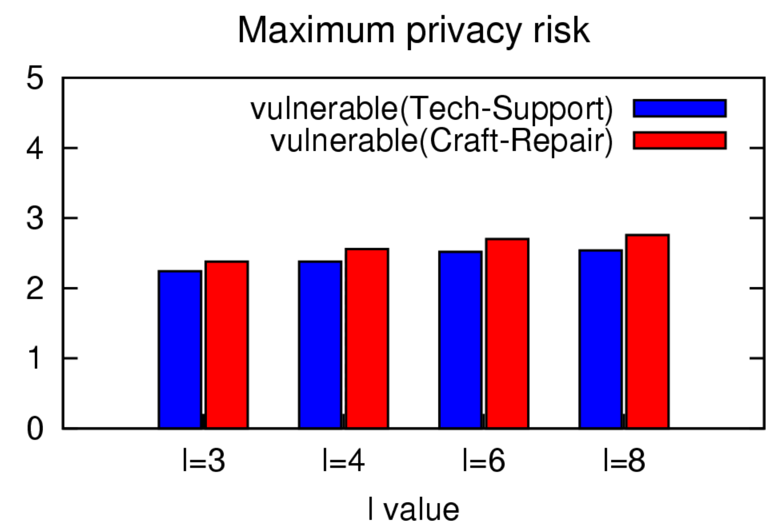
- So applying the minimality attack on this algorithm increases probability from $1/l$ to at most e/l
  - In first bucket, probability is at least $2/l$

Using bound on $p(b_l)$

- Simply set l based on e/l probability
- Apply randomization
  - Inference was possible due to predictability of merging
  - Instead, randomly choose to keep going even when diverse
  - Higher probability of merging decreases $p(b_1)$

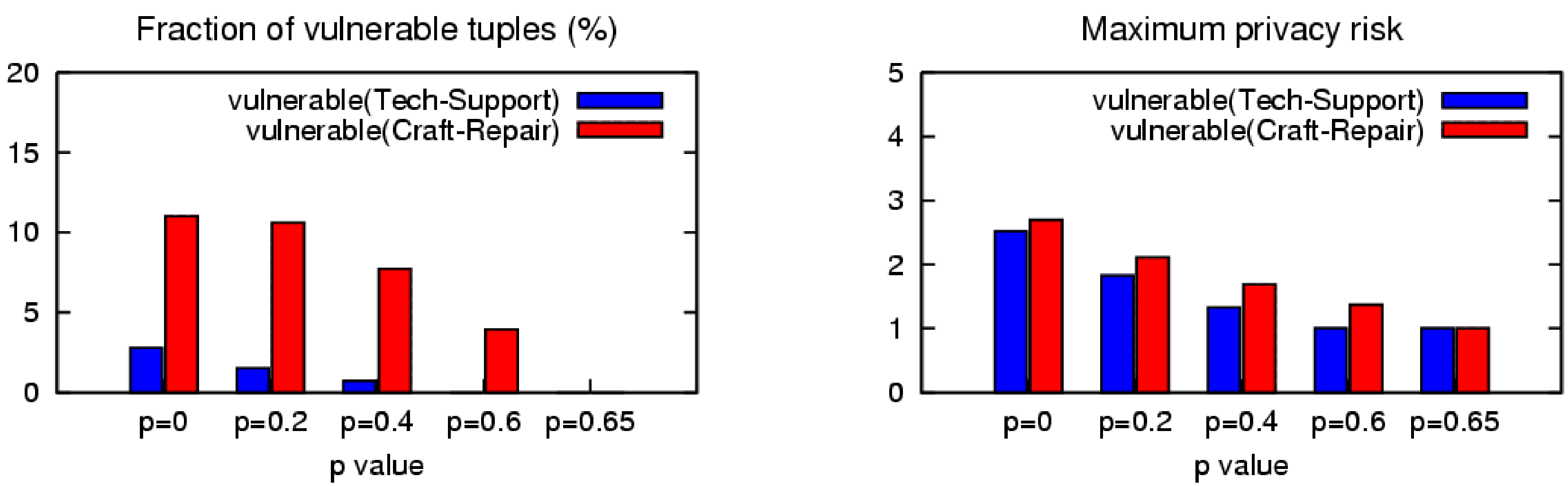
Experimental Study





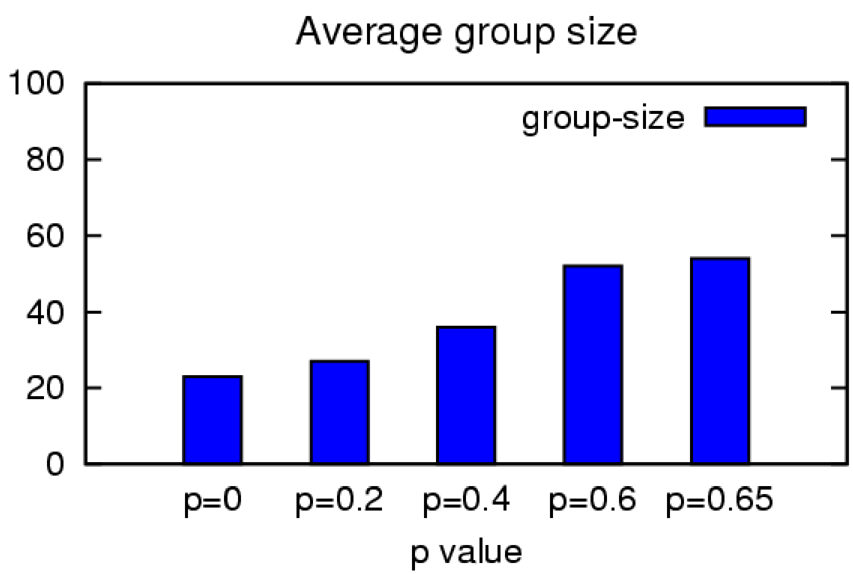
- Ran GG on UCI machine learning data set, career as SA
- Small fraction of tuples are vulnerable to attack
- Privacy risk as factor increase in probability < 2.7818

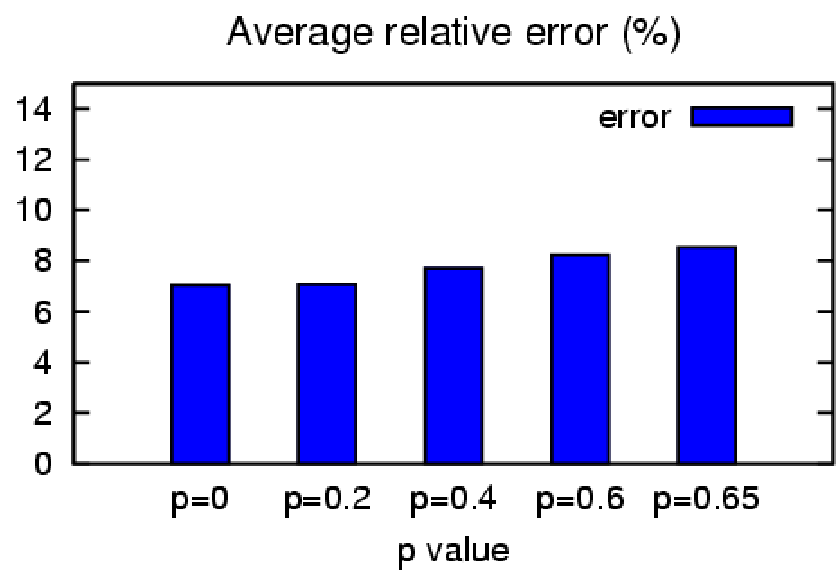
Use of Randomization



- Set l=6, randomly merge safe groups with probability p
- Rapid decrease in number of vulnerable tuples as p increases
- Privacy risk decreases to 1 (no increase in probability)

Utility Study





- ♦ Group size increases somewhat as p increases
- ♦ But accuracy of query answering barely affected!

Concluding Remarks

- Like crypto, anonymization proceeds by proposing anonymization methods and attacks upon them
  - Difference: Successful attacks on crypto reveal messages
  - Attacks on anonymization increase probability of inference
- Shown impact of minimality attack can be bounded
- Other attacks continue to be proposed
  - Use of inferred relationships to increase probabilities [Kifer 09]
- Long-term goal: propose anonymization methods which resist feasible attacks