Trading Privacy for Information Loss in the Blink of an Eye

A. Pilalidou *

FMT Worldwide

Limassol, Cyprus

P. Vassiliadis

Dept. of Computer Science,

Univ. Ioannina, Hellas

*work conducted in
Univ. Ioannina

Univ. of Ioannina

Summary

- Private data publishing involves hiding the relationship of a person with sensitive data for this person (typically via noise injection, or via hiding a person's info in a crowd of similar tuples). SoA suggests that a data curator anonymizes data **off-line**, by **trying to maximize the value of a utility function.** What if we refute this assumption?
- In this paper, we provide a method that allows the curator to **negotiate** information loss to privacy. We want to allow the curator to **explore different alternatives** in an attempt to reach an equilibrium on the **trade-off** of privacy relaxation vs. info loss (either via deleting outlier tuples or via abstracting more)
- To support this interaction, we (have to) provide :
  - **Instant** answers
  - **Recommendations** on alternatives
  - **Intuition** on decisions

k-anonymity

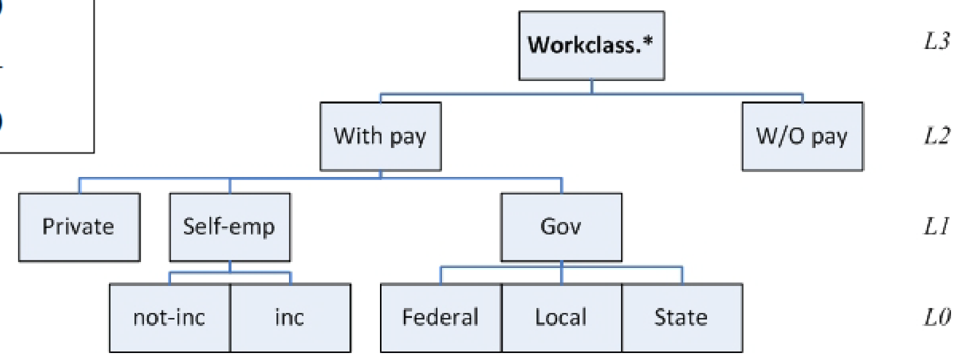
Name	Age	Work_class	Education	Hours/week
Thales	39	Private	Hs-grad	40
Anaximander	38	Private	Hs-grad	50
Anaximenes	37	Private	Hs-grad	40
Pythagoras	38	Private	11th	45
Gorgias	28	Loc-gov	Bachelors	30
Heraclitus	31	Federal-gov	Master	50
Empedocles	30	State-gov	Bachelors	60
Leucippus	32	Self-emp-not-inc	Bachelors	50
Democritus	35	Self-emp-inc	Prof-school	54
Protagoras	33	Self-emp-inc	Assoc-acd	40

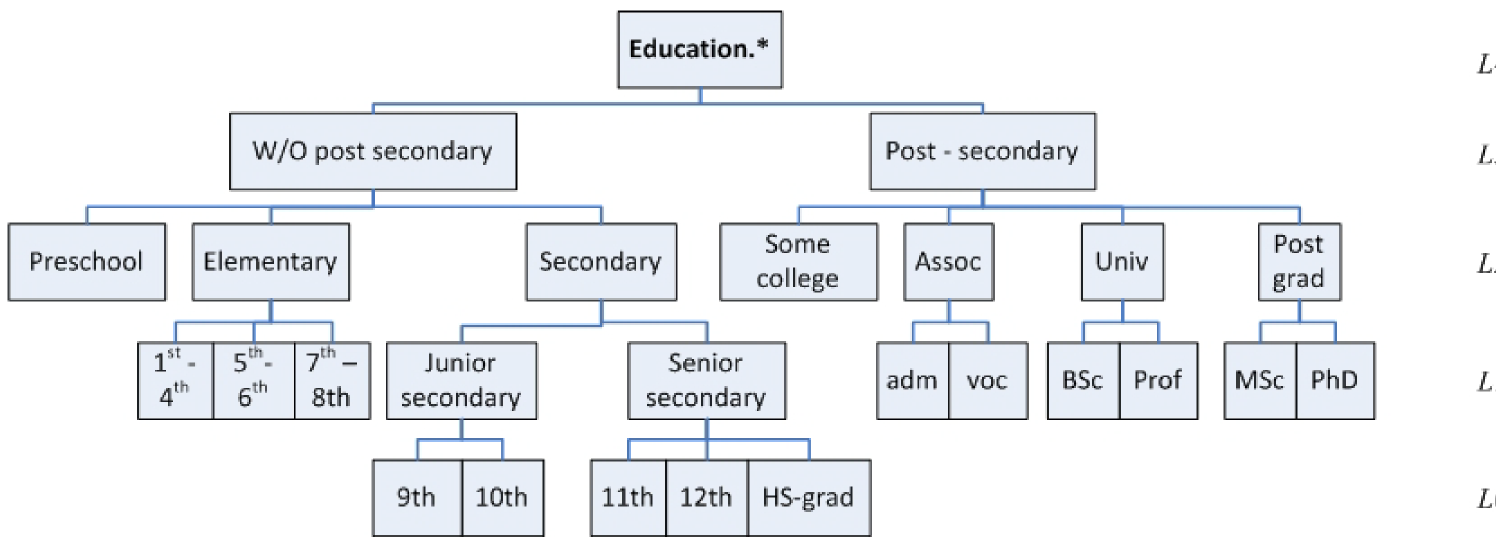
Name	Age	Work_class	Education	Hours/week
Thales	37-41	Private	Without-post-secondary	40
Anaximander	37-41	Private	Without-post-secondary	50
Anaximenes	37-41	Private	Without-post-secondary	40
Pythagoras	37-41	Private	Without-post-secondary	45
Gorgias	27-31	Gov	Post-secondary	30
Heraclitus	27-31	Gov	Post-secondary	50
Empedocles	27-31	Gov	Post-secondary	60
Leucippus	*32-36*	*Self-emp*	*Post-secondary*	*50*
Democritus	*32-36*	*Self-emp*	*Post-secondary*	*54*
Protagoras	*32-36*	*Self-emp*	*Post-secondary*	*40*

A relation *T* is ***k*-anonymous** when every tuple of the relation is identical to k-1 other tuples with respect to their *Quasi-Identifier* set of attributes.

Name	Age	Work_class	Education	Hours/week
Thales	39	Private	Hs-grad	40
Anaximander	38	Private	Hs-grad	50
Anaximenes	37	Private	Hs-grad	40
Pythagoras	38	Private	11th	45
Gorgias	28	Loc-gov	Bachelors	30
Heraclitus	31	Federal-gov	Master	50
Empedocles	30	State-gov	Bachelors	60
Leucippus	32	Self-emp-not-inc	Bachelors	50
Democritus	35	Self-emp-inc	Prof-school	54
Protagoras	33	Self-emp-inc	Assoc-acd	40

Hierarchies for the QI attributes allow the generalization of QI values and the formation of groups

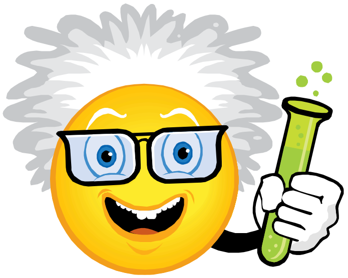




State-of-the-art

- All the related bibliography is based on the assumption that we have plenty of off-line time to process the data set
- The emphasis has been placed
  – To different privacy criteria and the corresponding attacks they prevent
  – To fast algorithms for exact solutions to the problem of optimal anonymization (wrt to a utility function)
    - Still: not fast enough for user-time (in the order of minutes / hours / …)

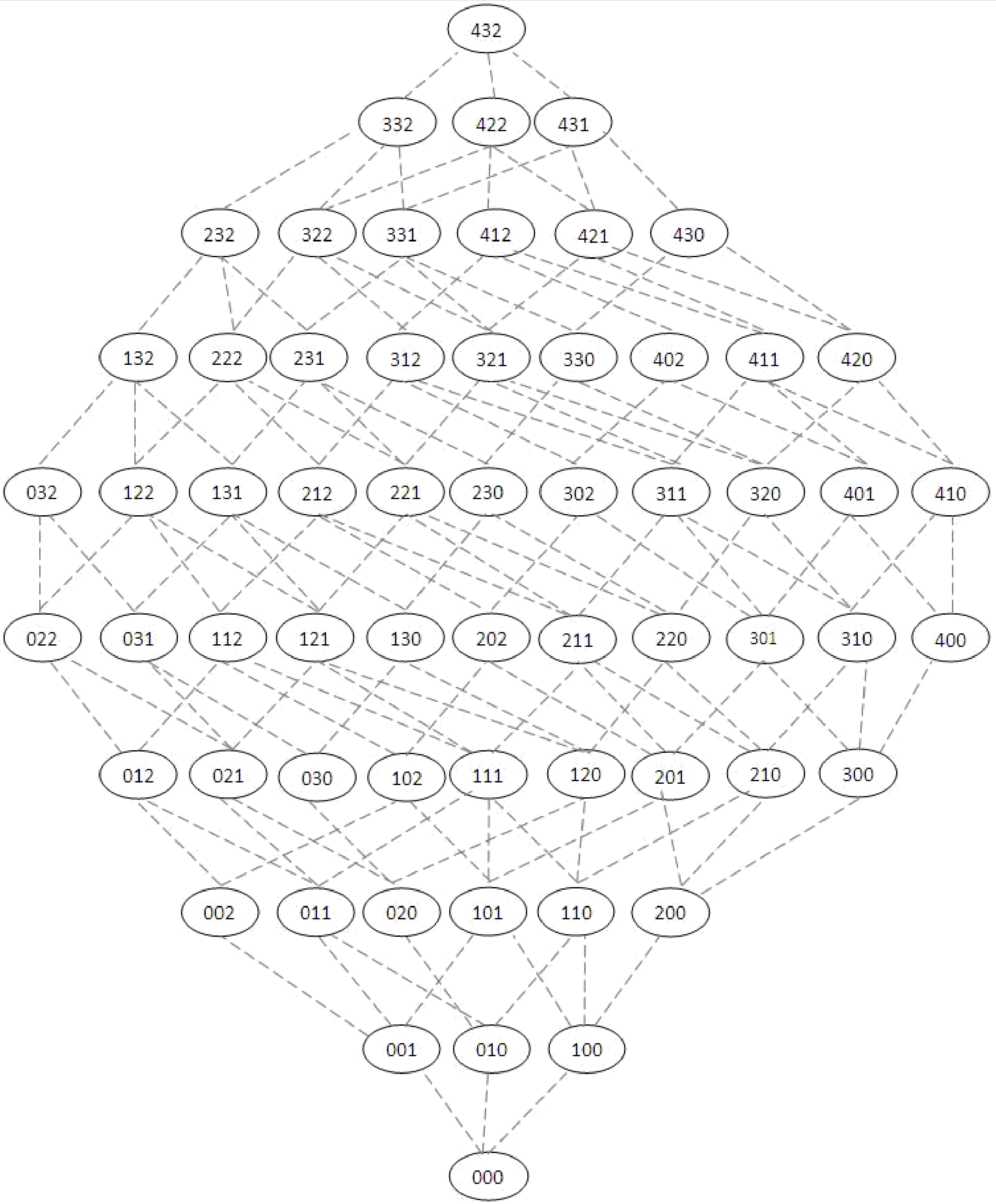
Research questions



- Can we help the data curator negotiate different configurations of privacy, generalization and suppression and decide what is best without resorting to some non-intuitive utility function?
  - e.g., by paying the price for less privacy (lower k) to attain a better value of suppression (less removed tuples) and, thus, higher information utility?
- Can the system guide the search by suggesting alternatives – esp., when tested configurations are impossible to attain?
- Can we do it in user time (i.e., without delays noticeable by the user)?

Our method

- We pre-compute, off-line
  - All the possible combinations of levels for the QI attributes – organized in a lattice of anonymization schemes
  - The suppression histogram of each such combination (for a specific privacy criterion) – i.e., for every combination we know the amount of tuples that have to be suppressed for a specific value of the privacy criterion
- The user specifies a request with 3 parameters as constraints (max height per hierarchy, max tolerable suppression, min tolerable *k* or *l*).
  - If a solution for this value combination exists
    - Among all the solutions that satisfy the request, we present the solution that is located at the lowest generalization height
  - If no such solution exists
    - we provide the user with 3 suggestions (i.e., approximate answers), each relaxing one of the 3 abovementioned constraints



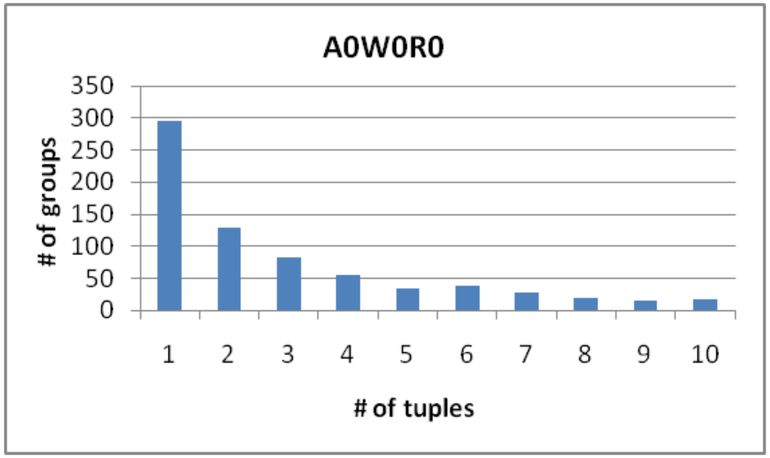
The anonymization lattice

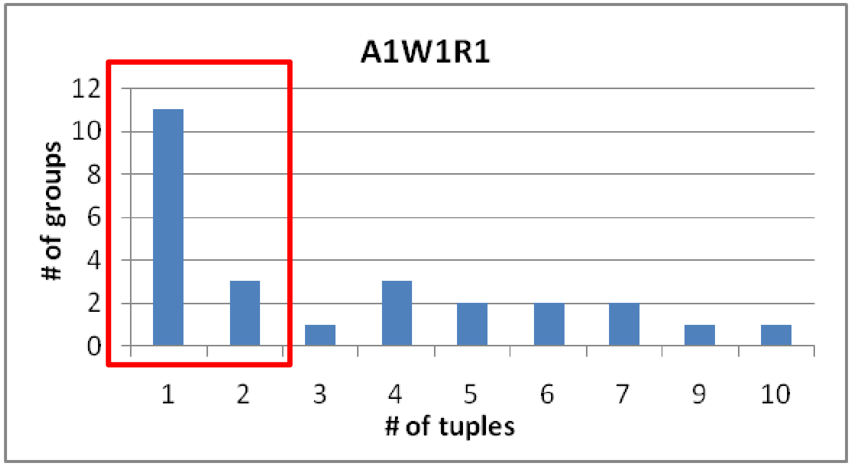
Here, QI=3 (Age, Workclass, Education, each with its own hierarchy)

A node is annotated by the levels of its QI attributes

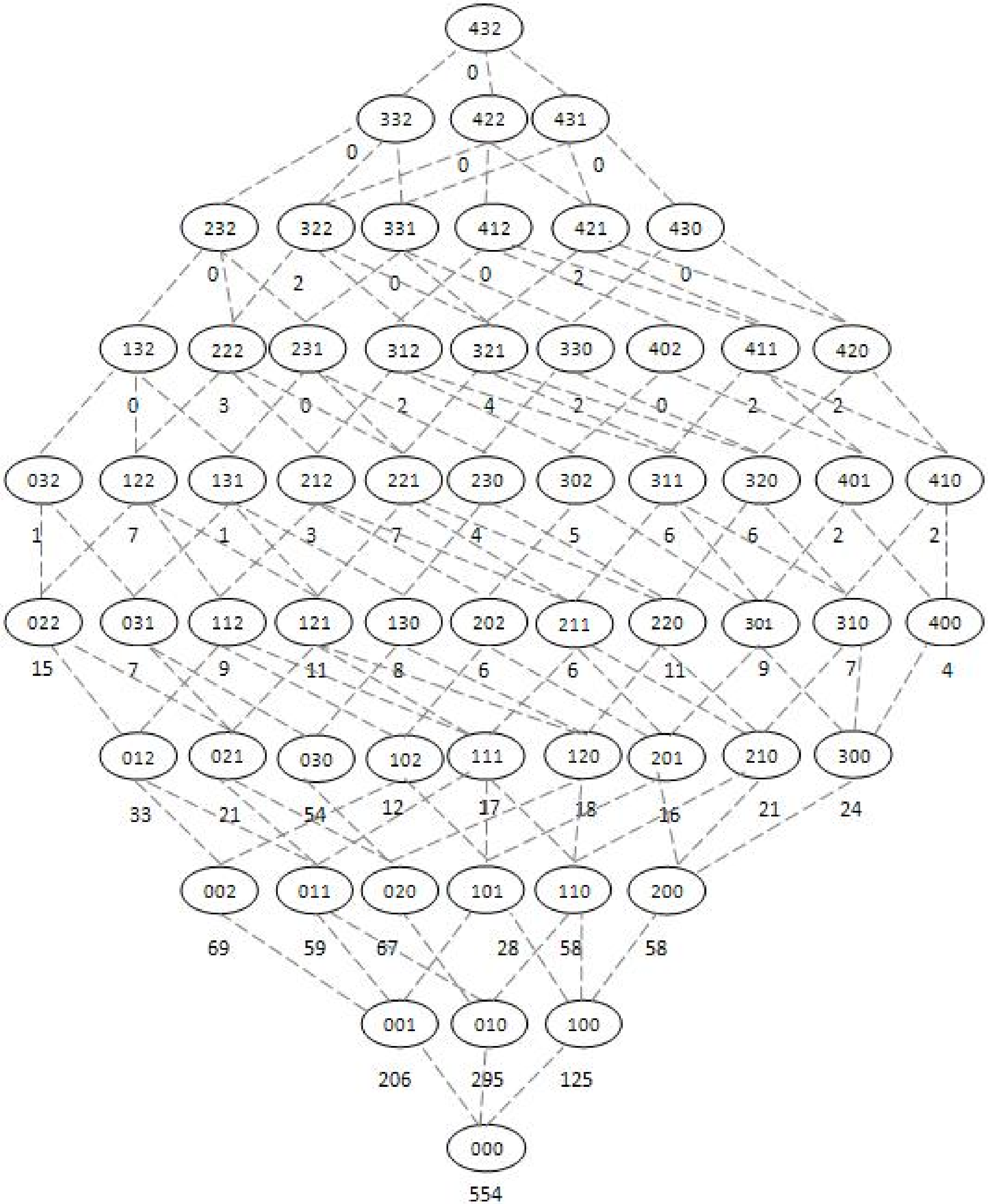
Eg. 302 means
L3 for age
L0 for workclass
L2 for education

Histograms

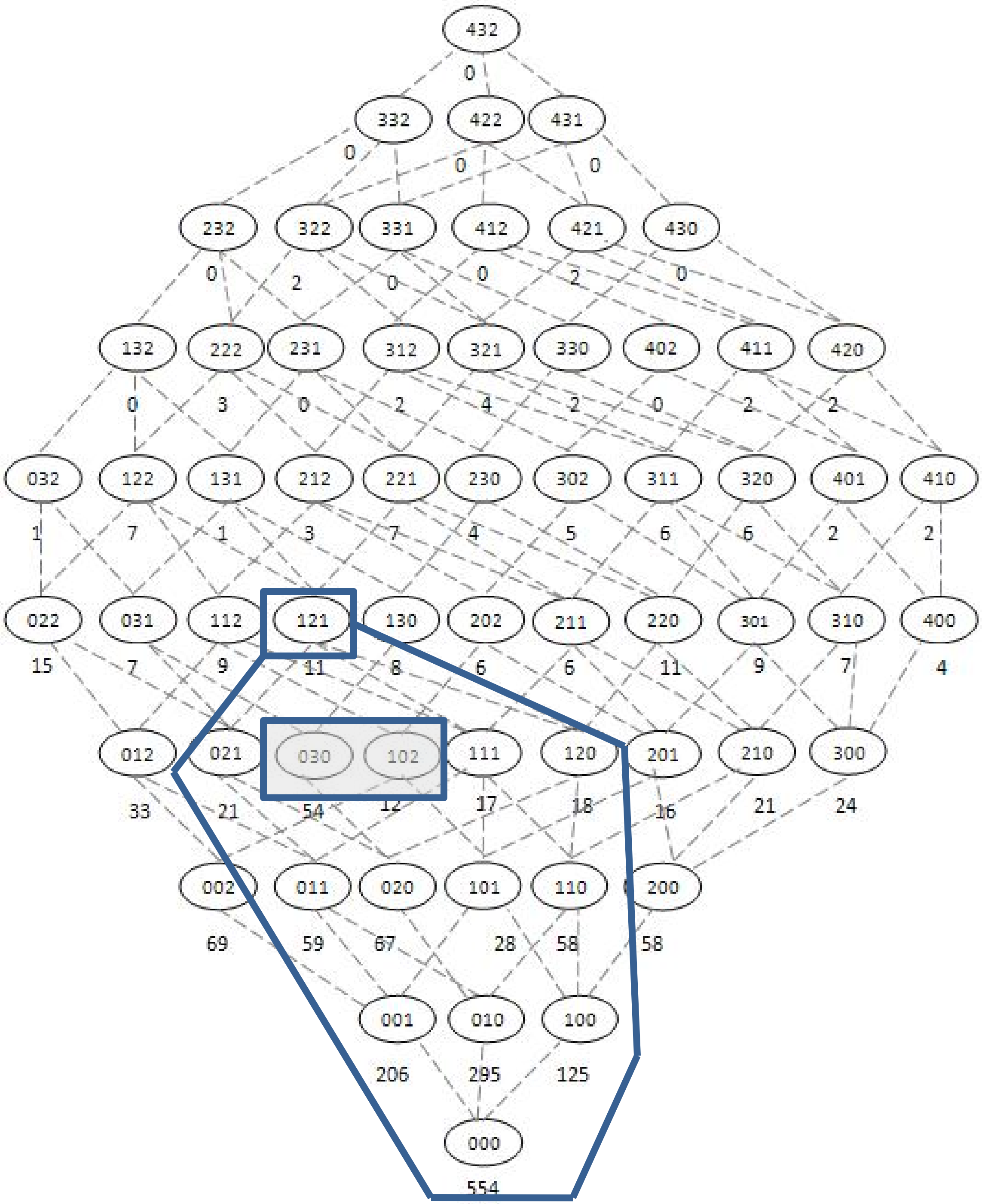




- Histograms allow us to compute the amount of suppression for a given value of k (equiv. l).
- E.g., to achieve 3-anonymity in level A1W1R1 we must suppress groups with size 1 or 2 => 17 tuples (17=1*11+2*3).1580 (1*834+2*373).

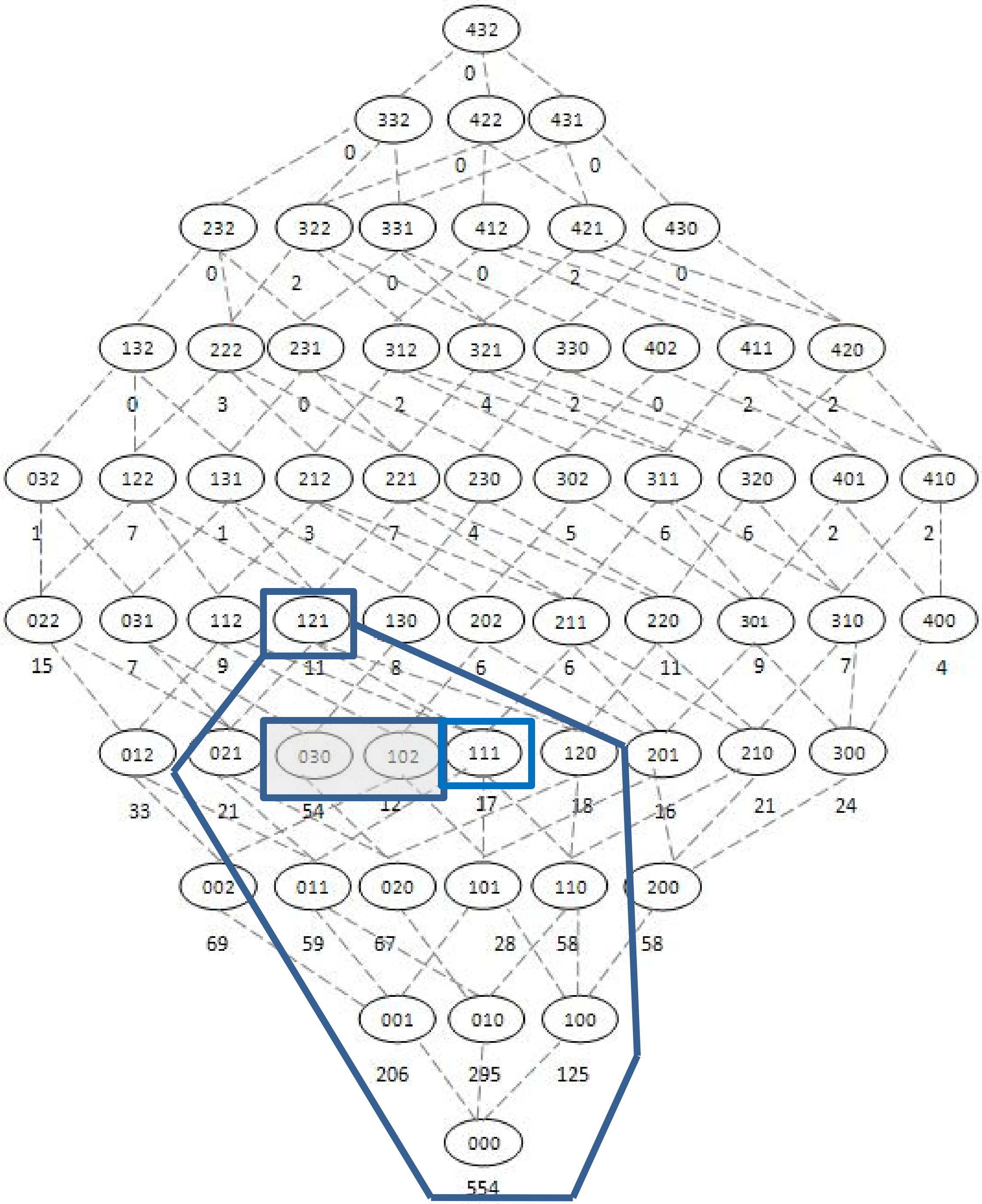


This is the lattice for QI=3 annotated with the number of suppressed tuples for k=3



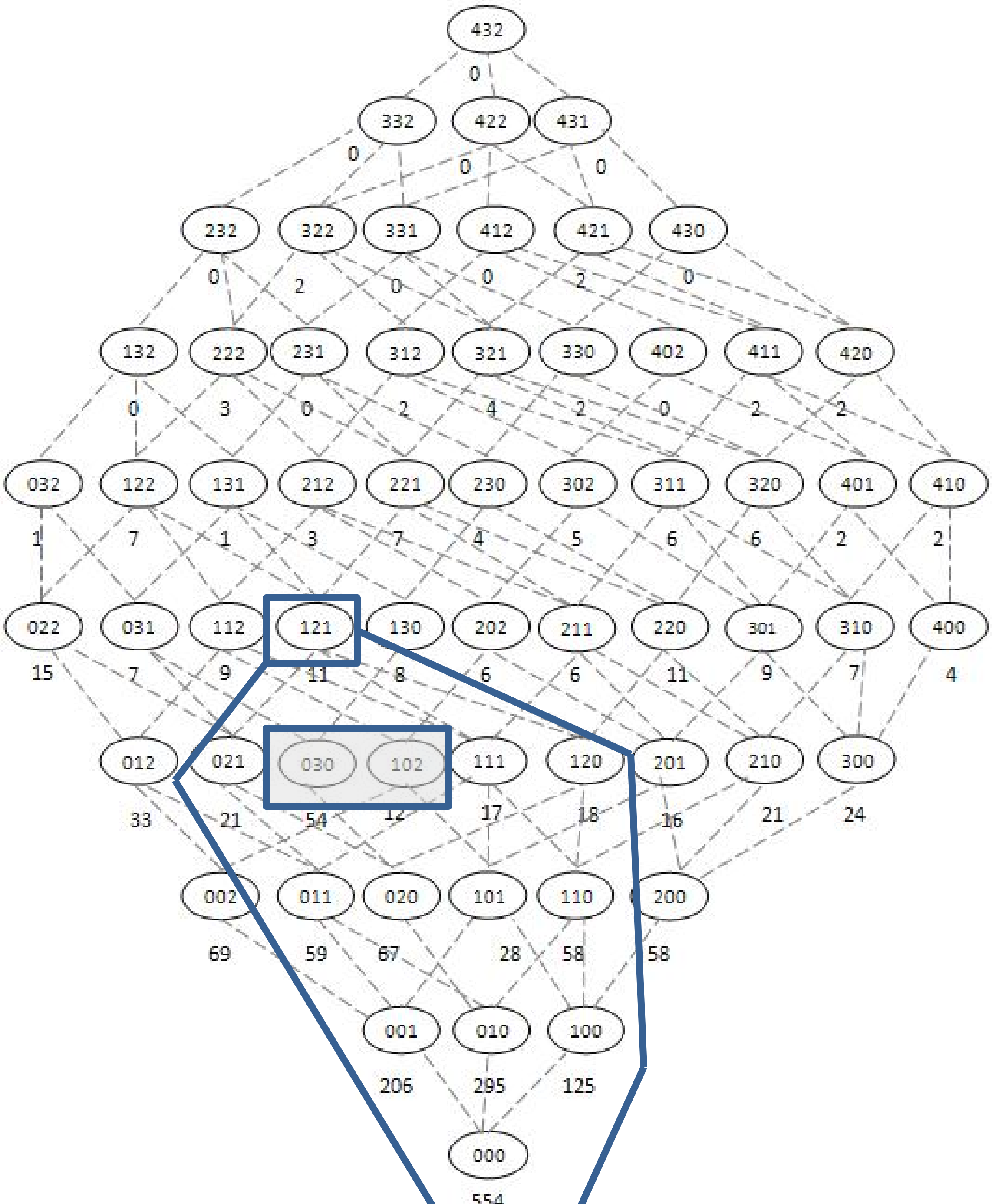
Assume the user requests:
h = 121
K = 3
MaxSupp = 20

Observe vmax 121



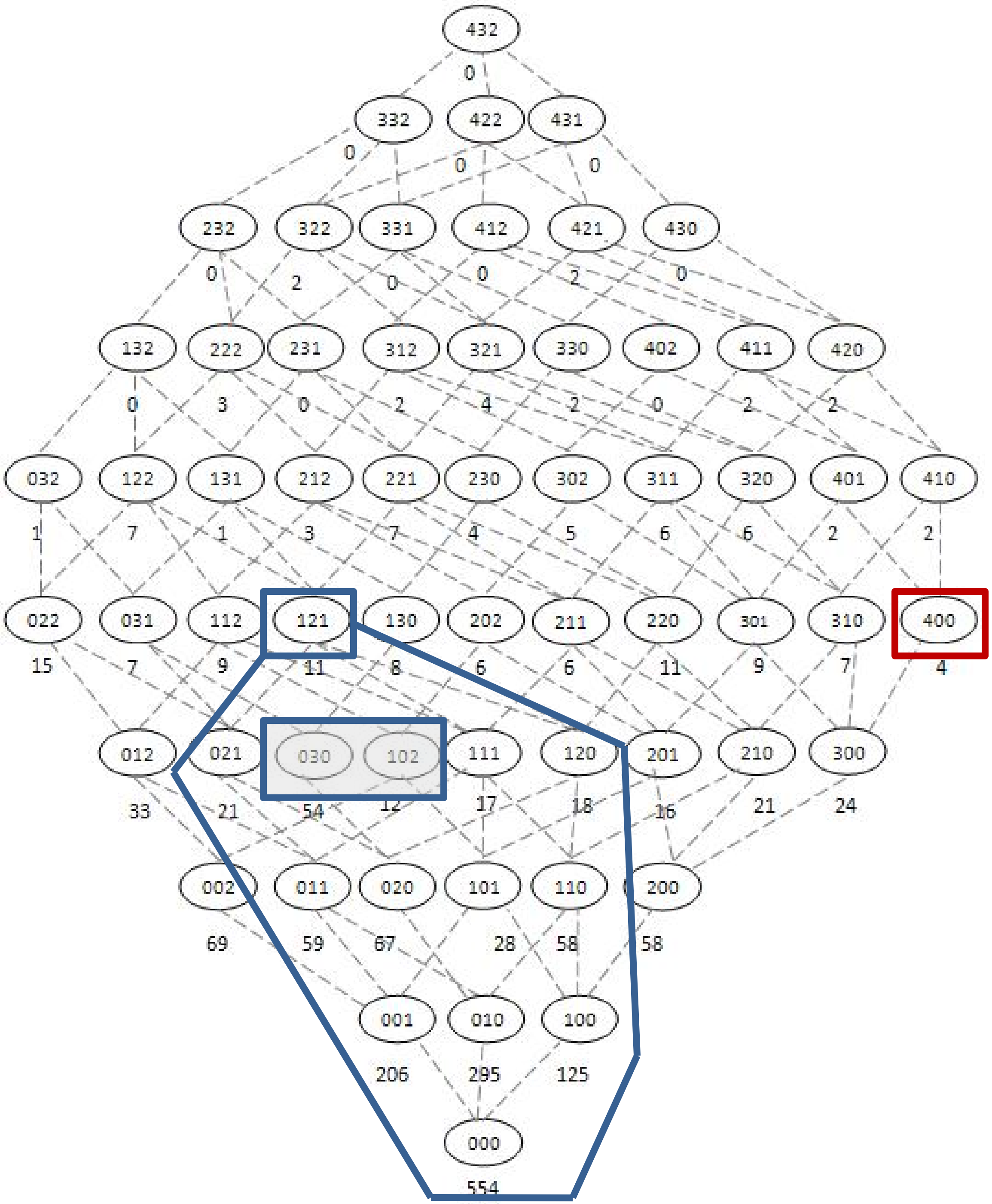
Assume the user requests:
h = 121
K = 3
MaxSupp = 20

The exact solution is 111 with #supp.=17



Assume the user requests:
h = 121
K = 3
MaxSupp = 8

Observe vmax 121: it fails to meet all three constraints



Assume the user requests:

h = 121

K = 3

MaxSupp = 8

Suggestions:

<u>Closest k:</u>

Node 121, k=2

<u>Closest height:</u>

Node 400, h=4

<u>Closest maxSupp:</u>

Node 121, maxsupp=11

Algorithm at a glance

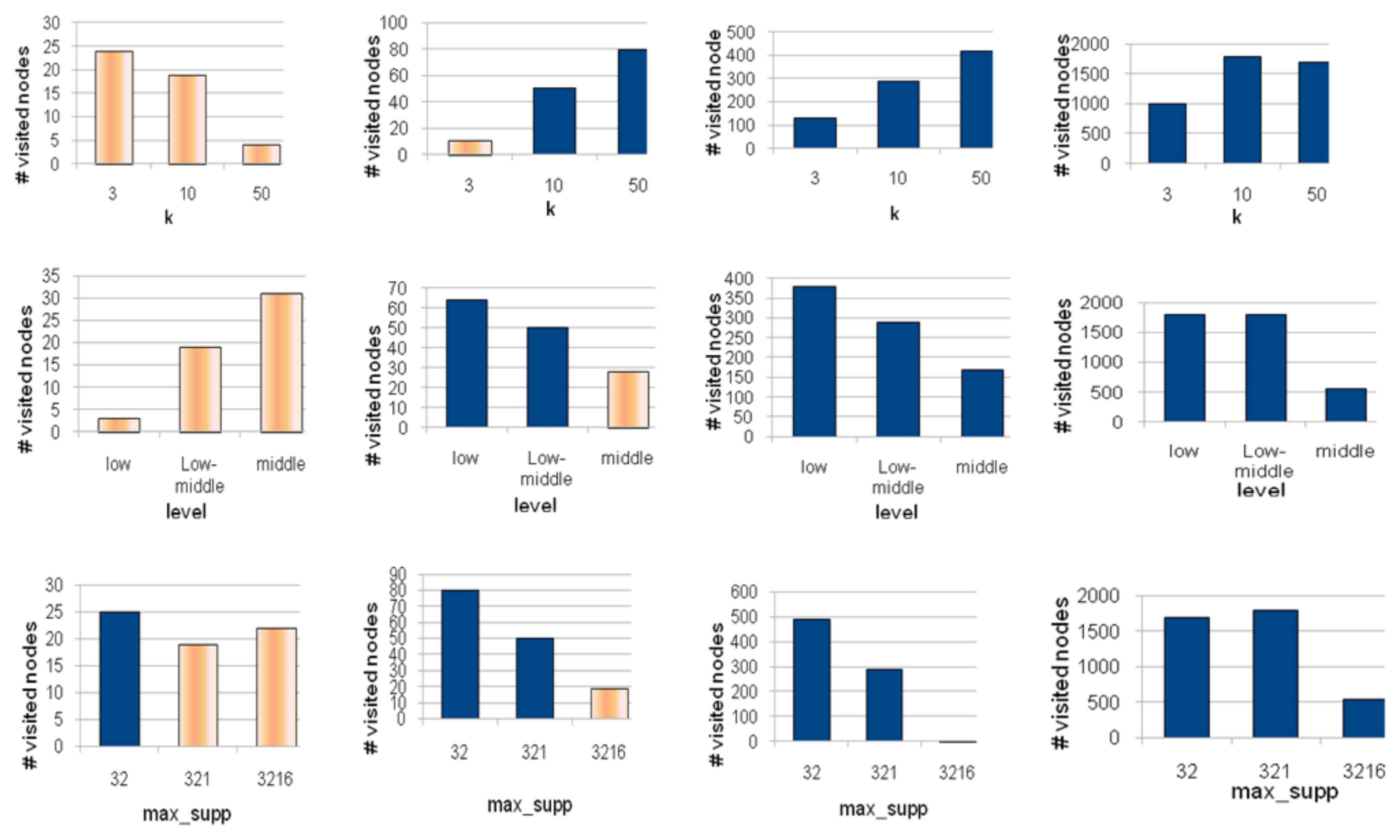
Input:

- Input relation R + hierarchies **H** + lattice with histograms
- A user request (k, **h**, maxSupp) with the user constraints

1. Identify top-acceptable node v_{max}
2. If v_{max} answers the (k, **h**, maxSupp)
   - Search within the sublattice of vmax for the lowest possible node that also answers (k, **h**, maxSupp)
3. Else
   - Relax MaxSupp: stay at v_{max} (respect **h**) and find the suppression value for k (respect k)
   - Relax k: stay at v_{max} (respect **h**) and find the largest k that suppresses less than maxSupp (respect maxSupp)
   - Relax **h** (retain k, maxSupp) and answer outside the sublattice:
     - Binary search between v_{max} and lattice's top
     - Exhaust all nodes of a level: if nobody answers, binary search between top and this level; else, whenever a node answers, perform binary downwards
     - Stop when it is impossible to descend and the last level is exhaustively tested

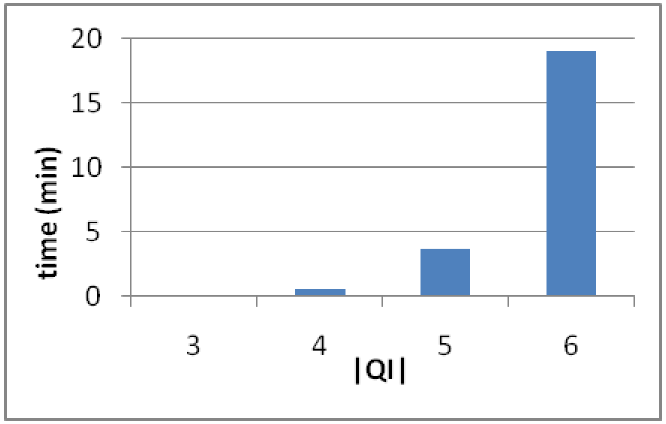
Some experimental results

Variant	QI = 3	QI = 4	QI = 5	QI = 6
k				
h				
MaxSupp				

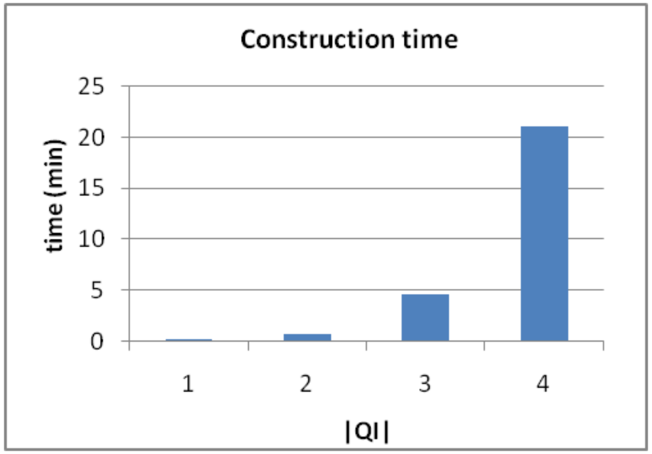


Number of visited nodes for different QI, k, h, MaxSupp. All times range between 1 and 8 msec. Light coloring is for exact matches and dark coloring is for approximate matches

Histogram construction time



K-anonymity



Naïve l-diversity

- Clearly dependent upon QI size, with an exponential tendency
- Remember that this is a compute-once use-many situation

To dig deeper …

- Can we respond in user time to anonymization requests? Can we suggest anonymization schemes that are approximately close to the original user request?
  - Yes to both! We have two ways to address the above, depending on the price we are willing to pay wrt the offline preprocessing of the lattice
  - Full lattice (preprocessing & query answering)
    - Exact answers and suggestions in less than 10msec (depends upon lattice size)
    - 18 sec – 20 min preprocessing time (depending on both the QI and the data size)

- The **long v. of the paper** (also long v of the talk) contains:
  - Theoretical guarantees that our method is guaranteed to provide the best possible answers for the given user requests.
  - Extensive discussion on the validity of the problem. To the best of our knowledge, this the first systematic study on the interdependency of suppression, generalization and privacy in a quantitative fashion.
  - Extensive experimental results, over the IPUMS and the Adult data set.
  - Partial lattice: o handle the issue of scale (as the off-line lattice-and-histogram construction is dominated by both the QI size and the data size) we provide a method for the selection of a small subset of characteristic nodes of the lattice to be annotated with histograms, based on a small number of tests that rank QI levels for the grouping power.

http://www.cs.uoi.gr/~pvassil/publications/2012_SSDBM/index.html

Thank you!

Questions?

A great many thanks to everybody who helped organizing SSDBM’12!