Diophantine inferences from statistical aggregates on few-valued attributes

Neil C. Rowe

Department of Computer Science
Code 52
Naval Postgraduate School
Monterey, CA 93943

Research on protection of statistical databases from revelation of private or sensitive information [Denning, 1982, ch. 6] has rarely examined situations where domain-dependent structure exists for a data attribute such that only a very few independent variables can characterize it. Such circumstances can lead to Diophantine (that is, integer-solution) equations whose solution can lead to surprising or compromising inferences on quite large data populations. In many cases the Diophantine equations are linear, allowing efficient algorithmic solution. Probabilistic models can also be used to rank solutions by reasonability, further pruning the search space. Unfortunately, it is difficult to protect against this form of data compromise, and all countermeasures have disadvantages.

1. Two problems

Consider a university personnel database, and the set of salaries of faculty. Suppose there are only three ranks (assistant, associate, and full professor) with salary the same for all members of a rank. Suppose we know from reading the catalogue the number of faculty at each rank, and suppose we know from the annual financial report the total amount of salary paid to professors (or equivalently, the mean for the institution). Then we can write a linear Diophantine (or integer-solution) equation in three variables, and solve for the the salary associated with each rank. We will in general obtain a finite set of possible values for each salary, which can be pruned if we know additional information such as reasonable limits on faculty salaries or the restriction they be multiples of one thousand dollars.

There is a kind of dual problem to this one. Suppose we know the total tonnage of British naval vessels in the South Atlantic, and suppose we also know from published sources the tonnages of the only types of ships owned by the British navy. We can then write a linear Diophantine equation and solve for all possible fleet compositions. If we know information such as bounds on the total number of ships in the British fleet we can narrow the possibilities further.

These two examples represent what we call the "unknown-values" and "unknown-counts" Diophantine problems, respectively. (Or "Diophantine compromises", but the latter word is mostly used for individual data-item value revelations, and the inferences here are about sets, though occasionally sets of size 1.) They arise whenever the following conditions all hold in regard to some attribute A and some set S:

1. A is numeric

2. A has only a small number of distinct values for the set S

3. we know the sum of all the values of A for S (or equivalently, the mean of the values and the size of S)

4. we either know (a) the number of items having each value of A for this set (the unknown-values problem), or (b) the exact values that do occur for A for this set (the unknown-counts)

5. the unknowns are drawn from a finite universe having not "too many" members.

By "small" in (2) we mean on the order of 10 or less, and by "too many" in (5) we mean on the order of 1000 or more. Situations that satisfy these restrictions arise usually with "artificial" attributes representing invented codes and measured properties of man-made objects. Often they involve joins, either explicit or implicit, of a small relation with unique-valued attributes representing fixed properties of objects with a larger relation representing relationships or activities of those objects.

Mathematically:

$$N1 * V1 + N2 * V2 + N3 * V3 + \ldots = T$$

where the V's are the possible data values, the N's the number of occurrences of each, and T the total. If the N's are known and the V's are unknown, we have the unknown-values problem; if the V's are known and the N's are unknown, we have the unknown-counts. In order to ensure that this equation is Diophantine, since the V's may be rational, we should divide both sides by the greatest common divisor of the V's.

U.S. Government work. Not protected by
U.S. Copyright.

If rounding and/or truncation is used in calculating T so that the result is not exactly the sum (as may significantly occur with totals of large sets), we can often infer the true sum since the greatest common divisor of the V's divides the right side an integer number of times. We round T to the nearest integer multiple of that divisor.

2. Multiple constraints

Knowledge of the sum on some attribute is a linear equality constraint, using the teminology of optimization. There are many additional ways of obtaining both equality and inequality constraints on the Diophantine solution, making protective countermeasures for the database difficult. Of course, if we obtain as many independent equations as variables we can often determine them uniquely. But even when we have more unknowns than equations the Diophantine (integer) restriction itself can narrow the possibilities to a small number.

We can briefly summarize the categories of additional equality constraints that may be available (see [Rowe, in press] for more details of each except the last two):

1. Additional moments on the data (e.g. standard deviation), which give linear Diophantine equations for the unknown-counts case, nonlinear for the unknown-values. The sum of the values can be considered the zeroth moment.

2. Corresponding moments on attributes that are in one-to-one relationship to the attribute of interest (i.e. that show "extensional" functional dependencies in both directions, functional dependencies true only for a particular database state), and hence have the same frequency statistics. These give additional linear Diophantine equations.

3. A generalization of the preceding, corresponding statistics on attributes that have an extensional functional dependency in only one direction, to or from the attribute of interest. These give linear Diophantine equations on new variables that relate by sums to the old variables.

4. Different factorizations of the same data, as in multidimensional contingency tables of sums (these give linear equations)

5. Statistics on unions of sets of interest, whether directly or indirectly by inserts to the database (these give linear equations)

6. If type checking is not enforced, moment calculation routines can be applied to few-valued nonnumeric attributes, giving equations like those of item 3 (giving linear equations).

7. If transformations of data values (e.g. logarithm, square root, reciprocal) can be applied before computing moments, Diophantine equations of the same form but generally different coefficients are generated. These are linear for the unknown-counts problem, nonlinear for the unknown-values.

8. Even multi-argument arithmetic operations on data values can sometimes be exploited. For instance, knowing the possible values for two attributes allows calculation of the possible values of their product, which if the former are integers are not uniformly spaced. Another example is knowledge of the proportion of items having a certain property in a set, when the size of the set is not known. This is a "Diophantine approximation" problem where we must find an a and b such that a/b is closer than some small error E to some proportion p. All solutions form series such that if a/b is a solution, then so is k*a/k*b for any positive integer k. For a random proportion, the number of solution series follows a binomial distribution, with the simplest solution requiring on the average a denominator of the square root of 2/E.

9. Joins can be used to create few-valued attributes as mentioned in the last section, but joins can also be used to get many additional equality constraints on a set. If we know the mean of an attribute of a relation, we can compare it to the mean of the same attribute after the relation has been joined with another on some other join attribute(s). We can take subrelations of the second relation, or use different second relations, or join on different attributes, to get a variety of different equations. The equations are (surprisingly) linear for the unknown-values problem, but nonlinear for the unknown-counts problem.

Additional inequality constraints can also prune the solutions possibilities for a set of Diophantine equations (again see [Rowe, in press] for more details):

1. bounds on the frequency distribution of the values for the unknown-counts problem, such as the mode frequency or the frequency of the least common item

2. bounds on the values for the unknown-values problem, such as absolute maxima and minima that it is impossible for values to go beyond

3. medians and other order statistics on the set, which state how many items can lie in a certain value range (useful for both unknown-counts and unknown-values)

4. the number of items having certain values in any superset containing the set of interest (needed for both). Often we know the number of items having particular values in the entire database, and we can also sometimes perform an easier Diophantine analysis (because of number-theoretic peculiarities) on a superset.

3. Solving the equations

We can take all the constraints found by the methods of the last section and find a finite set of possible values for each variable by a variety of methods. Fortunately, most of the abovementioned equality constraints are linear, and there exist sophisticated methods covered in [Chou and Collins, 1982] for finding solution vectors of matched variable values for this information. We can then apply inequality constraints successively, filtering out those vectors inconsistent with them. The possible values of the kth variable are then just the possible kth vector components.

But our goal of finding all possible values for a variable is less general, and we can take some shortcuts in the above approach. In particular, we can filter out many possibilities a priori whenever we can apply either of these rules:

1. In the Diophantine equation C1 * X1 + C2 * X2 + C3 * X3 + ... = T, where the C's are constants and X's are unknowns, find the pair of relatively prime C's that have the smallest product, and call them CJ and CK. Then for any other term I, XI can take any integer value from 0 to (S - (CJ * CK) + CJ + CK - 1) / CI. (This says nothing about larger values for XI.) This follows from the number theory result that for N > A * B - A - B, there exist some integers X and Y such that A * X + B * Y = N.

2. In the Diophantine equation C1 * X1 + C2 * X2 + C3 * X3 + ... = T, if for some I, XJ is divisible by some integer F > 1 for all J not equal to I, then (CI * XI) mod F = T mod F for any solution value XI.

Various methods can be used for nonlinear Diophantine equations too [Mordell, 1969]. General algorithms do not exist, but there are many special-purpose tools (e.g. analysis in modular arithmetic). An exhaustive combinatorial search can be fallen back on, since one can almost always find absolute bounds on the integer unknowns.

Unfortunately, it is very difficult to analyze the expected and worst-case time complexities of solution methods for the various Diophantine problems. No satisfactory results appear to have been published (though [Chou and Collins, 1982] do obtain some results for space required for linear equations). This is a serious problem for protection research because Diophantine inferences vary considerably in effectiveness with small changes in the coefficients involved (since the inferences come from number theory, they can be very sensitive to the lower-order bits of constants). To return to our first example of the paper, if there are 13 assistant professors, 7 associate, and 20 full, and their salary sum is $1,200,000, and we assume possible salaries are multiples of $1000, then we can use rule 1 above and say that assistants can have any salary from 0 to (1200 - 140 + 20 + 7 - 1) / 13 = 83.5 thousands, associates (1200 - 260 + 13 + 20 - 1) / 7 = 139.2 thousands, fulls (1200 - 91 + 13 + 7 - 1) / 20 = 56.4 thousands. But if a new full professor is added to the faculty, then all of a sudden we can apply rule 2 above and find that (13 * X) mod 7 = (-X) mod 7 = 1200 mod 7 = 3, X mod 7 = 4, and hence the only values possible for assistant professor salaries are 4, 11, 18, 25, 32, 39, 46, 53, 60, 67, 74, and 81 thousands. Bounds can rule out possibilities, so if say we know the range is $24,000 to $31,000, we can infer a unique value of $25,000, whereas with those same bounds before the professor was added we had 8 solutions.

The fact that the number of solutions to a Diophantine problem can vary so widely (and even more so with nonlinear Diophantine problems) means that someone wishing to intentionally facilitate unwarranted inferences (or a user with insert capability wishing to compromise) could choose a set of values or counts, perhaps just by fudging of true values, that could help enormously. And note an "easy" set of values or counts remains "easy" for any right-hand side constant, though constraints affect the actual number of solutions. Unfortunately, while some "easy" sets of values are apparent on inspection, others are not.

4. Ranking solutions

We can often do more than just obtain possibilities consistent with constraints. We can rank possibilities by reasonableness, perhaps assigning probabilities to quantify it. For instance, if an attribute represents the number of children in an employee's family, the value 10 is possible, but unlikely; so a solution for the frequency of values in a subset that has half of the employees with 10 children is highly unlikely.

A good general-purpose way of obtaining ranks for the unknown-counts problem is possible when the number of items having each particular value is known for the database as a whole. Then the distribution of values in any subset of the database can be modelled by a multinomial distribution if we assume independence of value occurrence, with probabilities equal to frequencies of values in the full database. Of course, knowledge of a particular database domain may suggest other "nonlinear" ranking methods which may supersede this. For example, for our professor data we may think that the difference between full and associate is probably pretty close to the difference between associate and assistant, and unlikely to be three or four times, or one third or one fourth.

5. Multivariable-dependent attributes

Thusfar we have required attributes with relatively few distinct numeric values. There is an important generalization to attributes with perhaps many values, but values all determined by a few independent variables. Consider the salary policy for most employees of Stanford University, roughly modellable as a logarithm of years of service, starting from a certain "level". Thus there can be many different employee salaries, but they can be explained by one of ten or so values for "level" and a variable for number of years of service (i.e., there is a functional dependency from level and years to salary). We can write an equation:

$$LI * \log(KY) = T$$

which we can make Diophantine by dividing by the greatest common divisor of the left hand side. So if

we know the levels and years for any subset containing all levels, and its salary sum, we can solve for possibilities for the LI and K. If we know more subsets, we can narrow the possibilities further.

6. Countermeasures

Good countermeasures against these inferences are hard. All methods have serious drawbacks.

Protection by limiting statistics computed on a set of data is one possibility, but may require computationally expensive analysis, since nothing less than attempting to solve every possible Diophantine situation in advance will do. Fortunately, however, the checklist given in section 1 is not satisfied very often, primarily because few-valued numeric attributes are rare. But they do arise from time to time, and when they do, little short of comprehensive threat analysis will do. Note that query overlap controls used to protect against a variety of classical compromise methods [Denning, 1982, ch. 6] are useless here because strong inferences can be derived merely from different queries on the same set, or even sometimes a single query. Controls that suppress statistics on particularly small sets are some help, but since the power of Diophantine methods is highly sensitive to the lower-order bits of coefficients, this doesn't help very much.

A better form of protection seems to be perturbation of the data or query output, since this can affect low-order bits severely. The perturbations have to be random, or the user might be able to discover the perturbation by experiment and possibility elimination, and they have to be pseudo-random as opposed to truly random, or the user might zero in on true values by asking the same query repeatedly. The perturbations must also be sufficiently large that user cannot just specify a small range of true values for statistics, given the perturbed values, and intersect all the results obtained from solving separately a Diophantine equation set for each possibility. Some of these equation sets may be immediately ruled out as impossible (e.g. anything with $4x + 10y + 18z = 101$ because the sum of even numbers can't be an odd number), while the equation set corresponding to the true values of the means and moments is always guaranteed to have a solution, the solution corresponding to the true state of the world.

A certain amount of query output perturbation may occur automatically in a database system due to rounding and/or truncation of statistics calculated on large sets. However, this perturbation may be quite systematic as opposed to random, and is likely to vary considerably in magnitude with the sizes of the sets being analyzed as well as the values being analyzed, and thus rarely can offer certain protection.

A curious aspect of Diophantine inferences is that they can still be possible even when the data is encrypted and statistical aggregates cannot be calculated directly. Assuming the encryption is on each data attribute value and not on blocks of attributes, the one-to-one nature of the encryption process which is necessary for recoverability makes encrypted data functionally dependent in both directions on the original data; hence we can use the methods listed under equality constraint 2 of section 2, provided we know the mean of the encrypted values, as we may quite easily in a public-key system. The solution is only a frequency distribution, and does not give correspondences between encrypted values and true values, but often many of these can be identified by inspection and methods similar to the solution of simple substitution ciphers by tables of English letter frequencies.

7. Conclusion

Diophantine inferences can pose an important threat to the confidentiality of certain kinds of data. Their power in specific cases is difficult to categorize short of detailed number-theoretic analysis. Protection measures involving witholding of statistics are weak in effect, and protection involving data or answer perturbation seems to be the only real possibility, with its concomitant disadvantages of degrading statistic quality. As best as we can tell, however, no publishers of summary statistics have addressed this type of compromise, including census agencies which have been concerned with other types [Cox, 1980; Sande, 1983]. It seems important that they become concerned.

8. Acknowledgements

The work reported herein was partially supported by the Foundation Research Program of the Naval Postgraduate School with funds provided by the Chief of Naval Research. It was also partially supported by the Knowledge Base Management Systems Project at Stanford University under contract #N00039-82-G-0250 from the Defense Advanced Research Projects Agency of the United States Department of Defense. Thanks to Gio Wiederhold, Dorothy Denning, and reviewers.

9. References

[Chou and Collins, 1982] Tsu-Wu J. Chou and George E. Collins, Algorithms for the Solution of Systems of Linear Diophantine Equations. SIAM Journal of Computing, 11, 4, November 1982, 687-708.

[Cox, 1980] Lawrence H. Cox, Suppression Methodology and Statistical Disclosure Control. Journal of the American Statistical Association, 75, 370, June 1980, 377-385.

[Denning, 1982] Dorothy E. R. Denning, Cryptography and Data Security. Addison-Wesley, Reading MA, 1982.

[Mordell, 1969] L. J. Mordell, Diophantine Equations. Academic Press, New York, 1969.

[Rowe, in press] Neil C. Rowe, Diophantine Inferences on a Statistical Database. To appear in Information Processing Letters. An earlier draft appeared as part of report STAN-CS-82-948, Stanford Computer Science Department, October 1982.

[Sande, 1983] G. Sande, Automated Cell Suppression to Preserve Confidentiality of Business Statistics. Proceedings of the Second International Workshop on Statistical Database Management, September 1983, Lawrence Berkeley Laboratory publication LBL-16321, 346-354.