Does Differential Privacy Protect Terry Gross' Privacy?

Krish Muralidhar [1] and Rathindra Sarathy [2]

[1] University of Kentucky, Lexington KY 40506, USA
[2] Oklahoma State University, Stillwater OK 74078, USA
krishm@uky.edu, rathin.sarathy@okstate.edu

Abstract. The concept of differential privacy was motivated through the example of Terry Gross' height in Dwork (2006). In this paper, we show that when a procedure based on differential privacy is implemented, it neither protects Terry Gross' privacy nor does it provide meaningful responses to queries. We also provide an additional illustration using income data from the US Census. These illustrations raise serious questions regarding the efficacy of using differential privacy based masking mechanism for numerical data.

Keywords: Differential privacy, Laplace noise addition, Numerical data.

1 Introduction

The concept of differential privacy was motivated through the example of Terry Gross' height in Dwork (2006, page 2) as follows:

> Suppose one's exact height were considered a highly sensitive piece of information, and that revealing the exact height of an individual were a privacy breach. Assume that the database yields the average heights of women of different nationalities. An adversary who has access to the statistical database and the auxiliary information "Terry Gross is two inches shorter than the average Lithuanian woman" learns Terry Gross' height, while anyone learning only the auxiliary information, without access to the average heights, learns relatively little.

Dwork (2006) then goes on to provide describe differential privacy and the Laplace based noise addition method to achieve the same. Although never explicitly stated, Dwork (2006) leaves the impression that the Laplace based noise addition would protect Terry Gross. But we never actually see the implications of using Laplace based noise addition and the level of protection it offers Terry Gross. In this paper, we carry this illustration to its natural conclusion to see the impact of using Laplace based noise addition to mask queries relating to the height of Lithuanian women, the extent to which it protects Terry Gross, and the implications of this approach for simple queries. We also provide an additional illustration using the incomes of individuals in the United States. These illustrations show that, for numeric data, the utility of the responses from a masking mechanism based on differential privacy is less than desirable.

J. Domingo-Ferrer and E. Magkos (Eds.): PSD 2010, LNCS 6344, pp. 200–209, 2010

2 Implementing a Differential Privacy Based Procedure

In the context of output perturbation, differential privacy is a standard which requires that the response to any query should be indistinguishable in the presence or absence of a single observation. An alternative description of this requirement would be as follows. Consider two databases D1 and D2 that differ in a single element. The response to any query from D1 should be "indistinguishable" from the response to the same query from D2 in a probabilistic sense if the responses satisfy the following requirement:

$$\frac{P[\kappa_f(D_1)=R]}{P[\kappa_f(D_2)=R]} \leq e^{\varepsilon}. \tag{1}$$

R is the response to the query from the system through the masking mechanisms $\kappa_f(D1)$ and $\kappa_f(D2)$ from D_1 and D_2, respectively (assuming, without loss of generality, that the larger probability is always in the numerator).

Furthermore, if the above requirement can be satisfied in the presence/absence of the most influential observation for a particular query, then this requirement will also be satisfied for any other observation. The impact of the most influential observation for a given query (Δf) can be assessed as follows:

$$\Delta f = \text{Max}\|f(D_1) - f(D_2)\| \tag{2}$$

for all possible realizations of D_1 and D_2, and where $f(D_1)$ and $f(D_2)$ represent the true responses to the query from D_1 and D_2. Dwork (2006) shows that if the response to any query is provided as f(X) + Laplace(0, b) where $b = \Delta f/\varepsilon$ (where X represents a particular realization of the database and f(X) represents the true response to the query), then such a response would satisfy equation (1). It is important to note that since equation (1) must be satisfied in the presence or absence of any observation, the evaluation of Δf must consider all possible realizations of D_1 and D_2; not just a particular realization of a database. In this sense, Δf represents the global sensitivity of the query. Please refer to Dwork (2006) and other papers for a more complete description of differential privacy and Laplace noise addition. We only consider the original definition of differential privacy (Dwork 2006). We do not consider any relaxations, such as found in Nissim et al. (2007), since they do not satisfy the original e^{ε} differential privacy.

3 "Terry Gross is Two Inches Shorter Than the Average Lithuanian Woman"

Lithuania has a population of approximately 3,400,000 with approximately 1,800,000 women. The average height of Lithuanian women is 167.5 cm or 66". Assume that we are Statistics Lithuania who has decided to respond to all queries regarding the height of Lithuanian women using Laplace based noise addition so as to satisfy differential privacy. For the purposes of this illustration, let us assume that Statistics Lithuania considers the height of a woman to be extremely sensitive and has set $\varepsilon = 0.01$.

In implementing Laplace based noise addition, we have to compute the maximum possible difference (Δf) that might possibly occur between two databases that differ in exactly one record for a particular query. Consider the simple Sum query. Let x_{min} represent the smallest possible value for a particular variable and let x_{max} represent the largest possible value for the same variable. For the Sum query, the maximum difference that could occur between two databases that differ in exactly one record $\Delta f = |x_{max} - x_{min}|$. For the purposes of this illustration, we will set $x_{min} = 23$" (the shortest woman according to the Guinness World Records) and $x_{max} = 98$" (the tallest woman according to the Guinness World Records) resulting in the global sensitivity $\Delta f = 75$. Hence, for all Sum queries, the scale parameter (b) of the Laplace noise distribution is set as $\Delta f/\varepsilon = 75/0.01 = 7500$ with the variance of the noise distribution = $(2 \times 7500^2) = 112500000$. Note that the number of records included in a particular Sum query does not affect the parameters of the Laplace noise distribution. Thus, whether the Sum query involves a single record or the entire population, noise is generated from a Laplace(0,7500) distribution. The response to any sum query would equal (The true value of the query + Noise from Laplace(0,7500)).

The Mean query is a simple variation of the Sum query where Mean = Sum/n. We can view this from two different ways. First we can view the Response to the Mean query as simply being the (Response to the Sum query/n). Alternatively, the noise generated for the Mean query would have noise generated from Laplace(0,7500/n) where n represents the number of records in the query. Either approach would result in exactly the same response distribution for the Mean query.

Now consider the response to the query "What is the average height of Lithuanian women?" Suppose the database responds with the true average height of Lithuanian women, namely 66. In Dwork (2006) this information can be used by an intruder to compromise Teryy Gross' height because the intruder has the auxiliary information that Terry Gross is 2 inches shorter than the average Lithuanina woman. To prevent this, the database has implemented Laplace noise addition and the distribution of responses to this query is provided in Figure 1. This figure shows that, even based on an extremely high level of security ($\varepsilon = 0.01$), the response to this query will be within (66" + 0.03") with probability close to 1. Once this response is received, the intruder knows that Terry Gross' height is very close to 64". Thus*, even with a very high level of security, differential privacy offers very little protection to Terry Gross since the intruder is able to estimate her height within 0.03"*.

Unfortunately, although little protection is offered to Terry Gross, the noise addition mechanism has a negative impact on other queries. Consider for instance, the following query: "What is the average height of women living in Smalininkai, Lithuania?" The city of Smalininkai has a total population of 621 with (let us say) 350 women. With a query involving 350 women, one would expect a reasonable answer to be within say + 1". The probability of observing a response within + 1" of the true average height is approximately 5%. The probability of observing a response within + 6" of the true average height is approximately 24% and the probability of being within + 12" is only 43%. In other words, in 57% of the cases, the response from the system would be outside 12" of the true height of 350 women. Clearly, users would consider such a response to be of little or no value.

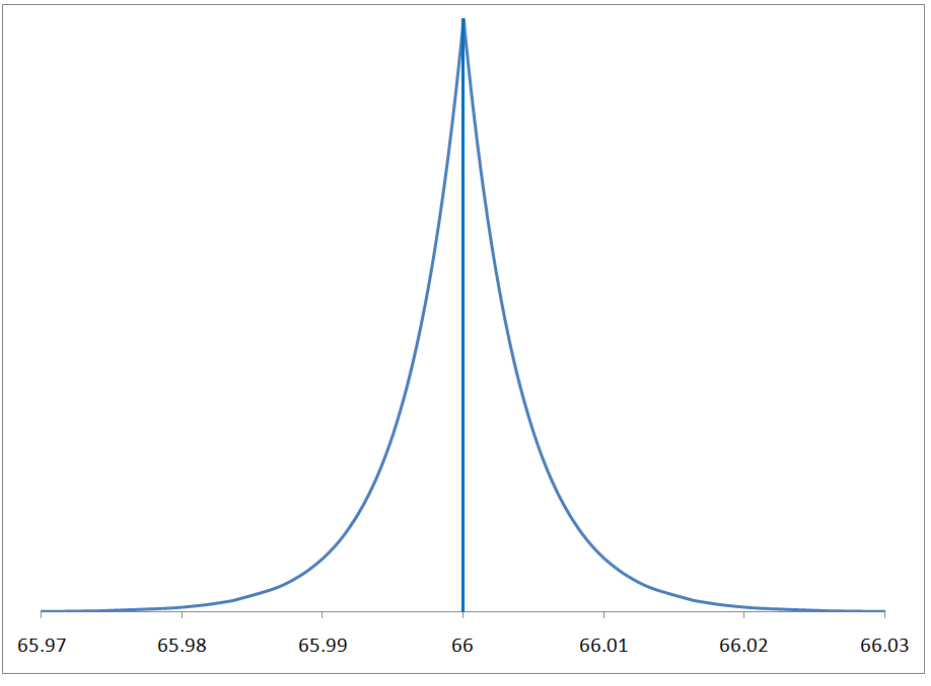

Fig. 1. Response distribution for the average height of Lithuanian women

Now consider the query, "What is the average height of all employed women living in Smalininkai, Lithuania?" Let us say that the number of women satisfying this query is 120. In this case only about 17% of the responses will be within 12" of the true value while the remaining 83% of the responses would fall outside the range. Consider the query "What is the average height of all employed women over age 50 living in Smalininkai, Lithuania?" Assume that the number of women who satisfy this query is 17. Only about 3% of the responses would fall within + 12" of the true value. Thus, in a vast majority of the cases for such queries, the response from the system would be practically useless. The probability of observing responses within specified limits is provided in Table 1 which clearly shows that for small subsets, Laplace based noise addition provides very little utility.

Another interesting aspect of Table 1 is the last column in the table which shows the percentage of cases where the response from the system is within the range of 23" to 98". Even for n = 350, 18% of the responses are either below 23" or greater than 98", which from a practical perspective makes no sense at all. With n = 120, a majority (55%) of the responses are not even within the (23" to 98") range. With n = 17, we get the ridiculous situation where 92% of the responses are outside this meaningful range. To illustrate this further, consider the distribution of the responses to the query "What is the average height of all employed women in Smalininkai?" is provided in Figure 2. For the purposes of this illustration, we have assumed the true height of the 120 women in Smalininkai who are employed to be 66". Note that for most real life

Table 1. An assessment of the utility of the responses

Group	n	Probability that the Response is within ± k" of the True Value							Probability that Response is Reasonable (between 23" and 98")
		0.10"	0.50"	1"	2"	4"	6"	12"	
All Lithuanian women	1800000	100%	100%	100%	100%	100%	100%	100%	100%
All women in Smalininkai	350	0%	2%	5%	9%	17%	24%	43%	82%
All employed women in Smalininkai	120	0%	1%	2%	3%	6%	9%	17%	45%
All employed women in Smalininkai over age 50	17	0%	0%	0%	0%	1%	1%	3%	8%

purposes, a subset of size 120 would be considered large and one would expect a reasonably close estimate of the true height.

The response distribution in Figure 2 shows the lack of utility from the Laplace based noise addition method. As observed earlier, a majority of the responses (55%) fail the simple common sense test since they are not even within the upper and lower limits for height of women. Approximately 17% or the responses will result in average height less than zero, which is clearly unacceptable.

There is a curious contradiction when using Laplace based noise addition to satisfy differential privacy. In order to satisfy differential privacy it is necessary for the variable to be bounded. Yet, when we implement Laplace based noise addition, the resulting responses are unbounded! Thus, we are left with the contradictory situation of having to make the assumption that height of women must be between (23" and 98"), but many of the responses are outside this range.

The irony is that the very high security specification was necessary to protect Terry Gross. Yet, the resulting procedure offers little security to Terry Gross, and the intruder is able to estimate Terry Gross' true height to within 0.03" accuracy. Unfortunately, the resulting response distribution is so poor that for almost, if not all, subsets, the responses have no practical utility as far as the user is concerned. Note that our illustration is a very reasonable one assuming a variable (height of women) that has a natural lower and upper bound. Furthermore, the bounds are reasonably close to the average height which, one would expect, would result in reasonable responses. This, however, is not the case; the responses for even what would be considered as large subsets by most practical standards, are of little value to the user.

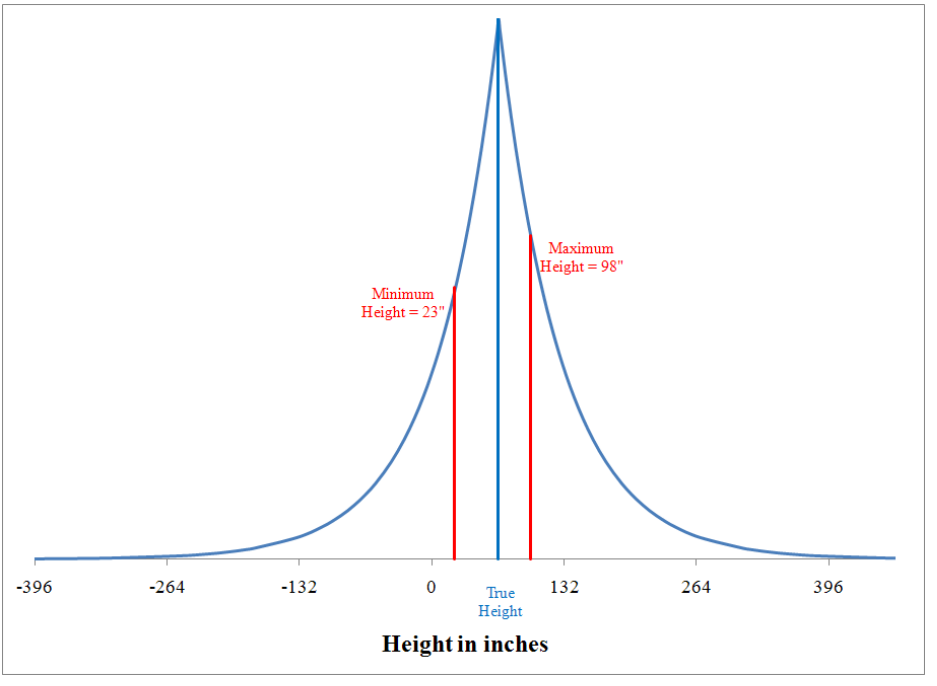

Fig. 2. Distribution of responses to the query "What is the average height of all employed women in Smalininkai?"

4 "Mr. Overy Rich's Income Is $5 Million More Than the Average American"

Consider the situation where the auxiliary information available is the following: "Mr. Overy Rich's income is $5 million more than the average American." Let us also assume that the variable Income is confidential. Let us also assume a security level of say $\varepsilon = 0.25$. It is well known that the income of some hedge fund managers exceed $1 billion[1]. In order to protect such individuals, it is necessary that Δf must be at least 1 billion. Note that, in order to satisfy differential privacy, it is better to be conservative in estimating Δf. For the purposes of this illustration, let us assume that Δf = 1,000,000,000. For this illustration, all information was gathered from the 2006-2008 American Community Survey at the U.S. Census Bureau web site.

There were a total of 97,488,418 individuals employed fulltime in the US with a mean income of $54,698. Assume that Laplace based noise addition is implemented to mask this data. Based on the specifications above, the range of the responses

[1] "James H. Simons, a former math professor who has made billions year after year for the hedge fund Renaissance Technologies, earned $2.5 billion running computer-driven trading strategies. John A. Paulson, who rode to riches by betting against the housing market, came in second with reported gains of $2 billion. And George Soros, also a perennial name on the rich list of secretive moneymakers, pulled in $1.1 billion." (Story, 2009)

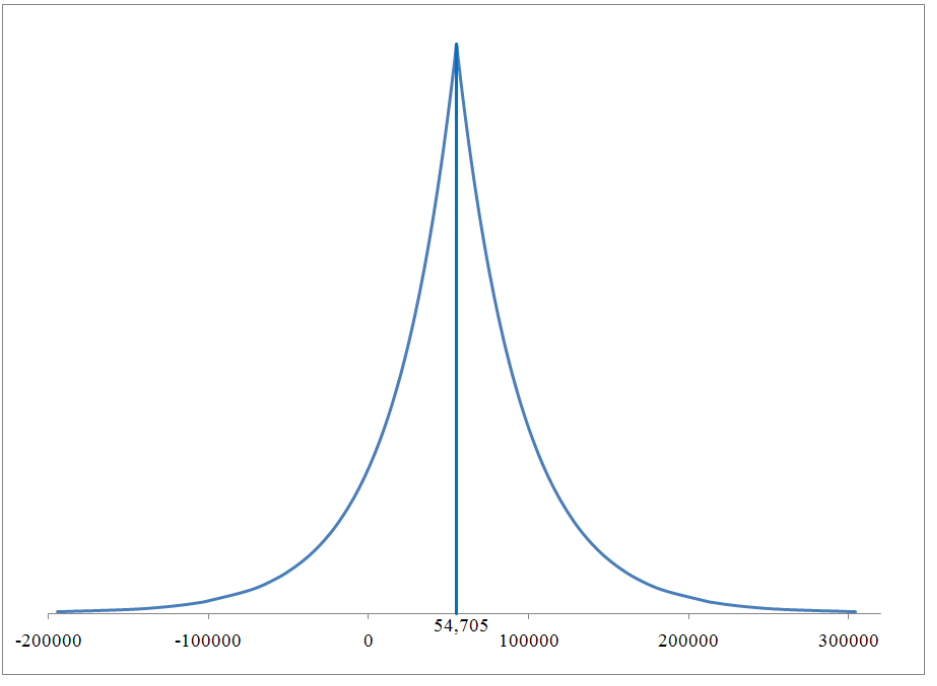

Fig. 3. Distribution of responses for the average income of Fayette County

from 0.1 percentile to the 99.9 percentile would be \$54,443 to \$54,953. In other words, in 99.8% of the cases, the response from the system would range between \$54,698 $\pm$ \$255. Hence an intruder would be able to estimate Mr. Overy Rich's true income as being between \$5,054,443 and \$5,054,953. From a practical perspective, this offers little protection to Mr. Overy Rich since we are able to estimate his income with a margin of error of about \$500 which is very small compared to his income. Thus, with the auxiliary information available to the intruder, differential privacy offers very little protection to Mr. Overy Rich.

By contrast, consider the legitimate query, "What is the average income of adults in Fayette County, Kentucky?" There were a total of 99,683 individuals employed fulltime in Fayette County with an average income of \$54,705. Assume that the response to this query is provided through the Laplace noise addition approach. In this case, the 0.1% of the response distribution would be –\$194,670 and 99.9% of the response distribution would be \$304,080. In other words, the range of the responses would be \$54,705 $\pm$ \$249,375. The distribution of the responses in this case is provided in Figure 3.

For a legitimate user, given that this subset consists of almost 100,000 individuals, it is reasonable to expect the response for the mean income to be very close to the true mean income. At most, one would expect a difference of say \$1000. But with Laplace noise addition, 98% of the responses would fall outside the $\pm$ 1000 range. In fact, 78% of the responses would fall outside the $\pm$ 10000 range. Approximately 12% of the

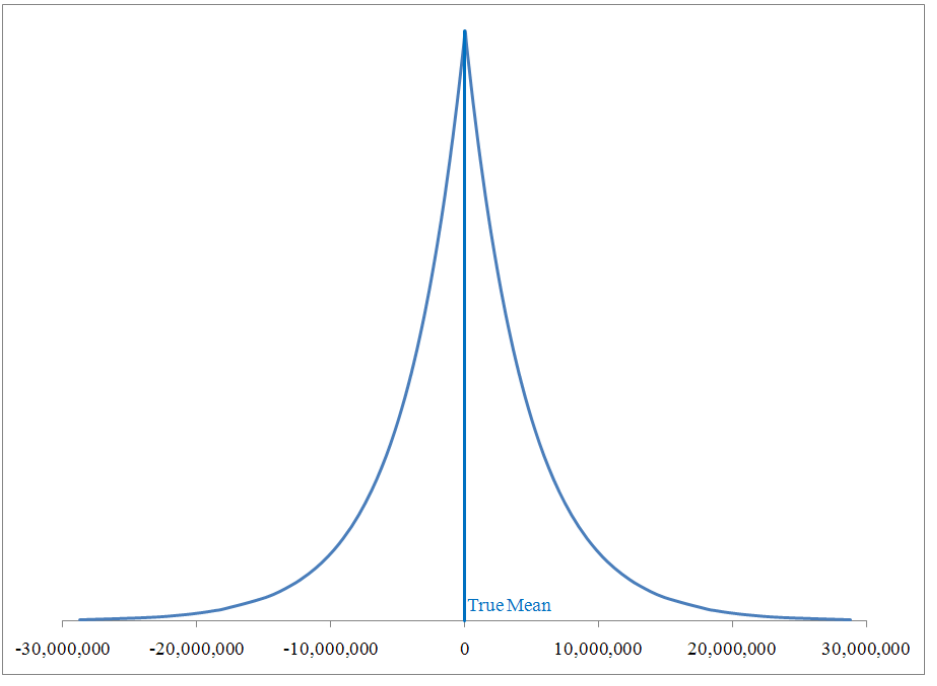

Fig. 4. Distribution of responses for the average income of women in Robertson County ($\varepsilon = 0.25$)

responses would result in average income less than zero. Approximately 17% of the responses would result in average income greater than $100,000, almost twice as much as the true average income. In essence, Laplace based noise addition offers little or no utility to the legitimate user even the size of the subset is relatively large (nearly 100,000).

Consider another similar query "What is the average income of adult women in Robertson County, Kentucky?" In this county, there were a total of about 863 women with an average income of $35,200. For this query, the (0.1%, 99.9%) range of responses would be between (– $28,769,472 and $28,839,872), which from the perspective of a legitimate user, is completely meaningless. Figure 4 provides the range of responses for this query. Only about 1% of the responses are within the range (0 to 100,000). Approximately 49.6% of the observations are below zero. Less than 1% of the responses are in the range $35,200 \pm 35,200$. These types of responses may be justified if the subset is very small. In this case, the size of the subset is, in practical terms, rather large (863). Even for a single record, it makes no sense to provide responses of the order of millions of dollars when the true value is only $35,200. In summary, given the true value is $35,200, when the vast majority responses are in the millions of dollars they offer no utility at all to the legitimate user.

For the purposes of illustration, let us consider the case where $\varepsilon = 10$. Note that this specification offers practically no security at all since $e^{\varepsilon} = e^{10} = 22026$. In other

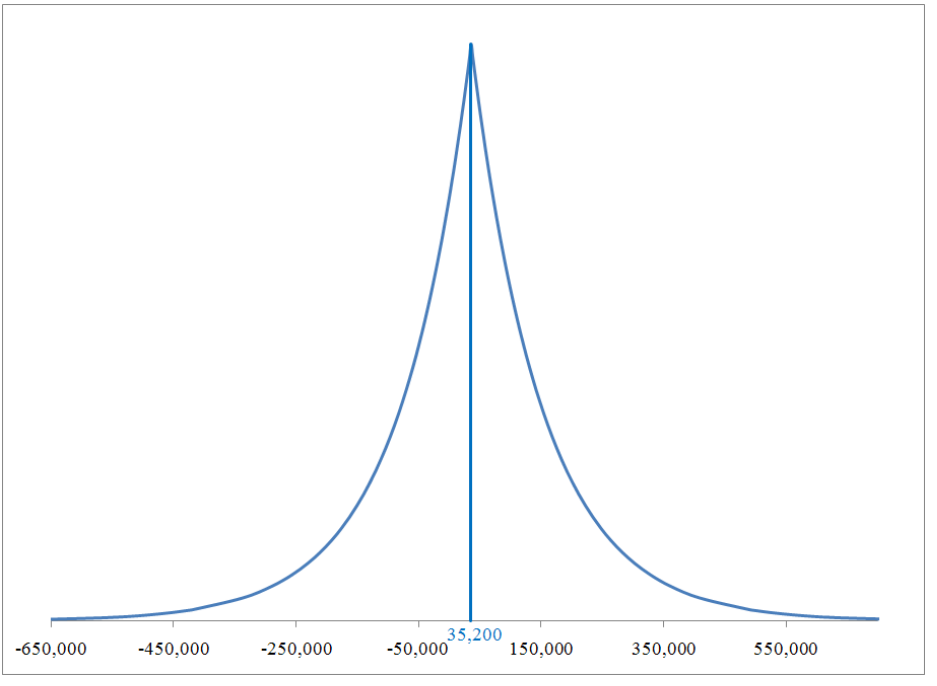

Fig. 5. Distribution of responses for the average income of women in Robertson County ($\varepsilon = 10$)

words, with this specification, the intruder's knowledge gain is of the order of 22026. The implication of this specification is simply that this specification offers practically no security at all. Unfortunately however, even though this specification offers no security, it does not improve the utility of the responses.

Figure 5 shows the distribution of responses for the average income of women in Robertson County with $\varepsilon = 10$. Note that the responses still range between (–658,000 to $755,000). Even with this low level of security, 35% of the responses are negative. More than 70% of the responses are outside the range 35,200 $\pm$ 35,200. Thus, even though ε is very large, the utility based on the responses is still very low and the entire procedure offers practically no security at all. Thus, we are left in the unenviable position of no security and not utility. This example also illustrates the fact that the value of ε makes little difference when Δf is large since the variance of the noise term will be dominated by the value of Δf. In these situations, the Laplace based noise addition will offer little or no utility regardless of the value of ε.

5 Conclusions

Differential privacy is being offered as a procedure for protecting privacy of records for all types of data. Unfortunately, the discussion on differential privacy is almost always limited to a theoretical discussion with a few minor exceptions where

illustrations for well behaved count data are provided (large cell count values, no sparse or empty cells, etc.). There are no practical examples of the application of Laplace based noise addition based on global sensitivity to satisfy differential privacy for numeric data. The real life numerical examples used in this study clearly show that the implementation of the Laplace based noise addition procedure in practice is likely to result in a situation where little disclosure protection is offered for large subsets and little utility is offered for small subsets. The behavior of the Laplace based noise addition procedure for numerical data also leads us to believe that in real life count data (with a large number of cells, sparse cells, zero valued cells, etc.), the Laplace based noise addition is likely to result in similar outcomes (where little security is provided for cells with large counts and little utility is provided for cells with small counts).

It is also important to note that these issues are the result of the inherent requirements of differential privacy. Differential privacy rests on this basic premise: "If I can make the two most extreme values for any query indistinguishable within a factor e^{ε} then all other values will also be indistinguishable." The problem with this approach is that when the magnitude of the difference between the two extreme values (global sensitivity Δf) is very large in relation to the variance of the dataset, the noise added is so large so as to make all responses to the query meaningless. This is clearly illustrated in our examples. Unfortunately, almost all economic data tend to heavily skewed and, in practice, it is likely that very large Δf values are the rule rather than the exception.

In summary, for numerical data, like Dalenius' (1977) definition of privacy before it, differential privacy is an interesting concept, but of little value in practice.

References

1. American Factfinder, U.S. Census Bureau, http://factfinder.census.gov/home/
2. Dalenius, T.: Towards a Methodology for Statistical Disclosure Control. Statistisktidskrift 5, 429–444 (1977)
3. Dwork, C.: Differential Privacy. In: Bugliesi, M., Preneel, B., Sassone, V., Wegener, I. (eds.) ICALP 2006. LNCS, vol. 4052, pp. 1–12. Springer, Heidelberg (2006)
4. Nissim, K., Raskhodnokova, S., Smith, A.: Smooth Sensitivity and Sampling in Private Data Analysis. In: Proceedings of the thirty-ninth annual ACM symposium on Theory of computing, pp. 75–84 (2007)
5. Story, L.: Top Hedge Fund Managers Do Well in a Down Year. New York Times, March 24 (2009), http://www.nytimes.com/2009/03/25/business/25hedge.html