Enabling Dynamic Linkage of Linguistic Census Data at Statistics Canada

Arnaud Casteigts[†‡*], Marie-Hélène Chomienne[‡Υ], Louise Bouchard[†‡], Guy-Vincent Jourdan[*]

[*]School of Information Technology and Engineering, University of Ottawa
{casteig, gvj}@site.uottawa.ca
[†]Institute of Population Health, University of Ottawa
louise.bouchard@uottawa.ca

[Υ]Dept. of Family Medicine, University of Ottawa
mh.chomienne@uottawa.ca

[‡]Institut de Recherche de l'Hôpital Montfort, Ottawa

Abstract—We explore the applicability of two privacy-preserving mechanisms to solve a real-world problem of strategical importance in population health research in Canada. The goal is to enable dynamic and *automated* linkage between health data (hosted at ICES – a *provincial* agency) and language variables from the 2006 census (hosted at Statistics Canada – a *federal* agency). The first mechanism relies on a pattern of interaction that, by design, prevents the collection of residual information, all the while making assumptions on the entities' intentions. The second makes no such assumption and relies instead on the true understanding of what statistical leakage is at play, thanks to recent advances in private data analysis (a field at the confluence of cryptology, statistics, and database systems). In particular, we show how results from this domain make it possible to characterize the worst-case information leakage, and support the argument that data utility remains achievable for virtually any level of risk acceptable. This demonstrates a form of linkage is possible between health data and linguistic data in Canada, despite the seemingly hard obstacles induced by its multi-level privacy policies. Based on these observations, we suggest some directions to implement the proposed mechanisms, whose availability would create unprecedented opportunities for population health researchers.

I. INTRODUCTION

A. The context

Research in population health consists of studying the impact of various factors (*determinants*) on health, with the long-term objective of yielding better policies, programs, and services. Determinants of health are many, and researchers of Official Language Minority Communities (OLMCs) focus specifically on those related to speaking a minority language, such as English in Quebec, or French in the rest of Canada. (Here, we will take Ontario as our focus.) Investigations of this type require, at the very minimum, the possibility of associating health data to linguistic information, whether at an individual or community level. Unfortunately, the largest health databases in Ontario, held at the Institute for Clinical Evaluative Sciences (ICES), do not contain usable linguistic variables to date.[1]

High-quality language variables from the 2006 Census exist at Statistics Canada (SC). Linking to the Census variables in the context of a punctual study is possible, and has already been done in Manitoba [1], but every such operation needs to satisfy a prescribed review and approval process [2]; it must also take place over a limited time, and requires to move the external data to be linked (health data, in our case) from its original source to SC, which may not be achievable repeatedly.

We are interested in finding ways to exploit linguistic data from the 2006 Census on a regular and automated basis, by enabling its linkage to ICES health data in a dynamic way. It must be clear that we do not consider here a traditional type of linkage, in which data from both parts are matched to produce a somewhat larger amount of information of an individual nature (whether nominatively or anonymously). The linkage we consider is intrinsically transient and aggregated: it consists in allowing ICES to learn interactively how many Francophones are present in a given sample of individuals. This simple operation, referred to as a *count query*, can reveal a powerful building block to answer more complex questions about OLMCs, as we will see. Count queries are actually a particular type of *tabulation*, an operation already practiced at SC but requires a manual process of

---

[1]Efforts are being carried out in this direction, as ICES is in the process of receiving linguistic information from Ontario's Ministry Of Health And Long-Term Care (MOHLTC). However, this information corresponds to the language individuals select for correspondence with the Ministry, which has potential for several biases and is not considered as reliable by OLMC researchers. We are currently running an independent project to assess the quality of this variable.

verification. Turning it into something automated poses a number of challenges, which are addressed in this paper.

We are concerned with ensuring privacy for both health and linguistic data. We suggest two possible mechanisms to enable dynamic count queries. The first consists of a circular workflow between the three involved entities: OLMC researchers, ICES, and SC. The workflow is initiated by the researcher through the submission of health criteria to ICES. A representative sample of individuals matching these criteria is then generated and sent to SC, which performs the count query. The result of the query is finally returned to the researcher. The privacy in this mechanism comes from the fact that the researcher does not know the sample details, SC does not know the health criteria that were used to generate that sample, and ICES does not know the final answer. This however assumes that no additional exchange of information occurs between the entities. In particular, the distinction between ICES and the researchers may be debatable from the point of view of SC (which generally considers anything from outside as one and a single entity).

The main problem when ICES and the researchers are seen as the same entity by SC is that the collection of residual information over several count queries becomes possible. Therefore, we propose a second mechanism that strives to prevent such a collection by adding noise to the queries answers. As it happens, this technique has seen a resurgence of interest lately [3, 4, 5, 6, 7, 8, 9, 10] and efforts have been devoted to understand the precise impact of noise on the privacy of the considered data. We adapt these results to characterize what leakage is precisely at play in our scenario, and how several parameters are involved in the tradeoff between leakage and utility. We rely on these results to argue that a safe exposition of linguistic census data could indeed be envisioned, and beyond, that similar techniques could be used to enrich provincial health databases in general with a vast range of federal census data, making it possible to perform fine-grained community-based studies in Canada.

B. Assumptions

For the sake of clarity, we will make three assumptions. These assumptions will be used recurrently in the document, and relaxed (or discussed) in Section IV. They are:

1) The language of an individual can be fully described by a Boolean value $\{Franco, Anglo\}$. (This over-simplistic view is erroneous, since language is a *multi-dimensional* variable.)
2) The census data comprises this variable relatively to every individual. (In fact, the unit of census is the *household*, not the individual.)

3) It is possible to generate, from the ICES database, a representative sample of individuals matching a given set of health criteria. (Whether this is possible is one of the questions we ask in this document.)

C. Count queries

Given a sample of individuals s, and a property p, a *count query* $q(s, p)$ consists in counting the number of individuals in s that satisfy p. We consider the following count query:
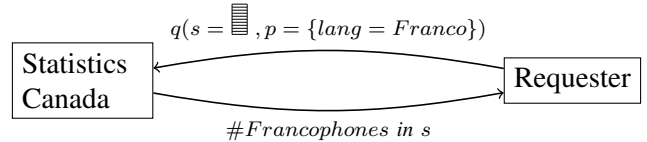

Fig. 1. A basic count query.

As such, a sample of individuals is sent to SC, and SC answers back with the number of Francophones in that sample. To illustrate the potential of this simple query, let us consider the following questions:

1) In Ontario, what is the average angioplasty rate among Francophones? (vs. Anglophones)
2) In Ontario, what is the hospital utilization rate for 65+ seniors with type-2 diabetes? (Francophones vs. Anglophones)

As far as *language* is concerned, the use of count queries can answer these questions. In the first case, ICES would generate a representative sample of individuals having undergone angioplasty, then ask SC for a count query in that sample. By normalizing the resulting answer over the more general ratio of Francophones in Ontario (which is known), one can answer the initial question. We can apply a variation of this method to the second example, by using several samples in various ranges of utilization rates by 65+ diabetic patients.

Hence, a single mechanism (count query) on a single variable (language) seems sufficient – as far as SC is concerned – to answer a variety of OLMC-related questions. Our concern is to integrate such a mechanism within ICES and use it in an automated way.

a) The problem: A malicious use of count queries, if unsupervised, could make it possible to identify the language of a given individual (say, Madame x). Consider the following attack: making a query with a sample s_1 that does not contain x; then making a second query with the same sample, plus x (see Figure 2). Obviously, if $answer(s_2) > answer(s_1)$, then Madame x is Francophone.

Even though language is supposedly not as 'sensitive' as other classes of information (e.g. health or income), we
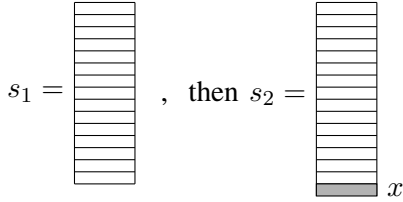
Fig. 2. Identifying the language of an individual by means of successive queries (the most basic example).

fully understand the necessity for SC to prevent leakages of this type, at least for the sake of it. Solutions to this problem are not trivial, even though one may think it suffices to prevent *this* particular scenario. A malicious adversary could actually build more complex constructs with similar effects.

We propose two solutions for the automation of language-based count queries at SC. The first avoids the problem by means of dataflow itself; while the second recognizes the risk and characterizes mathematically what tradeoff is at play between leakage and utility.

II. FIRST MECHANISM: TRIPARTITE INTERACTION

The first (and simplest) mechanism aims at providing guaranties at the data workflow level. It consists of a *tripartite* interaction between researchers, ICES, and SC; each entity obtaining and contributing only the minimal amount of information required to carry out the count query. The workflow is depicted on Figure 3, and can be summarized in three chronological steps:

1) Researchers issue a set of criteria related to a research question (e.g., angioplasty rates, or hospital utilization rate in a 65+ type 2 diabetes population);
2) ICES generates a sample of individuals responding to the criteria and sends it directly to SC;
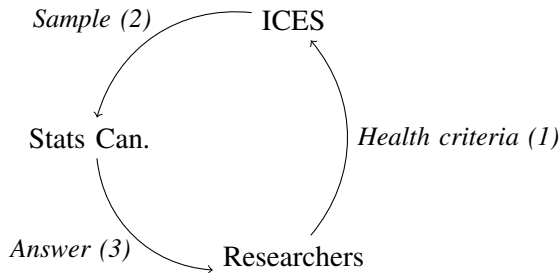3) SC executes the language query on the generated sample, and answers directly to the researchers.

Fig. 3. Privacy by means of tripartite interaction.

This scheme protects privacy through the combination of the three following statements:

1) SC does not know what health criteria the sample is associated with;

2) OLMC researchers do not know what sample resulted from the health criteria;
3) ICES does not know the result of the language query.

Protection of confidentiality of ICES health data results from statements 1 and 2, whose combination implies that ICES remains at all times the only entity capable of associating individuals with the queried health criteria. Similarly, the combination of statements 2 and 3 implies that SC remains the only entity associating a sample of individual with the number of Francophones in that sample; such an association is necessary to collect residual information (and thus perform attacks like that of Figure 2). However, these guaranties are subject to the assumption that entities will not collaborate beyond what is expected.

III. SECOND MECHANISM: NOISY COUNT QUERIES

This mechanism anticipates the rejection of the previous mechanism if SC considers ICES and the researchers as a single entity. The suggested idea is to add random noise in the query answer, in order to prevent the possibility of identifying an individual's language. Figure 4 illustrates this new interaction pattern.
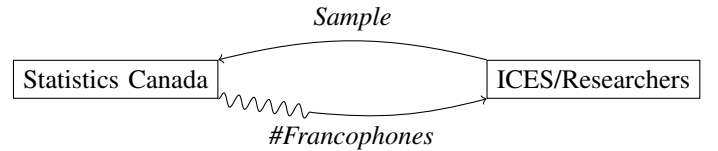

Fig. 4. Noisy count queries.

As we will see in this section, this mechanism involves the interplay between three parameters: 1) the magnitude of the noise, 2) the level of statistical leakage accepted by SC, and 3) the number of queries that can be performed (before shutting down the service). We will characterize this interplay mathematically, based on recent advances in cryptology, statistics, and database systems; then discuss some concrete options.

A. Overview of the principle

What makes possible an attack like that of Figure 2, and more generally the accumulation of residual information on individuals, is that the presence or absence of these individuals inside the sample does impact the answer. Suppose the exact answers for s_1 and s_2 are respectively 586 and 587 (*i.e.,* Madame x is Francophone) and consider adding or removing a small random number to the answer before returning it – we call it *perturbing* the answer with some *noise*. The resulting answers for

s_1 and s_2 may be for example 589 and 584, or 583 and 587, or 585 and 585. Clearly, these answers do not leak Madame x's language, and still, they are meaningful to a researcher.

There exist different types of noise, the most common of which are binomial, Gaussian, or Laplacian noises (named after the distributions from which the random number is drawn). As we will see, Laplacian noise has good properties that makes it an appropriate choice in our case. We consider perturbing the answer by adding a random number drawn from a Laplace distribution (see the example on Figure 5). This distribution can be considered at various *scales*, depending on how sharp or flat the probabilities are concentrated around the true result.

(a) Laplace distribution

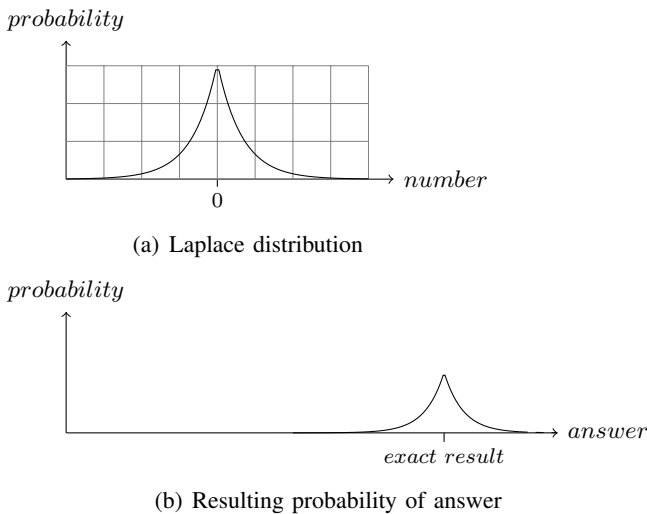(b) Resulting probability of answer

Fig. 5.   Addition of Laplacian noise.

Intuitively, performing a small number of *noisy queries* will not leak much individual information, whereas repeating a similar operation many times could eventually leak something substantial. In fact, *every query does leak a small amount of statistical information*, the quantity of which depends on the noise magnitude (scale of the Laplace distribution).

Consider again the example with s_1 and s_2, and two different scenarios, where noise is added according to a small scale or a large scale, respectively. The resulting probabilities of answers are represented on Figure 6. In a worst-case scenario, assume the requester already knows the correct answer for s_1, and is querying the system with s_2. It eventually obtains a random answer, say n.

Figure 6 shows intuitively how the leakage depends on the noise scale. On the left, with an arbitrary small scale, answer n is approximately 6 times more likely if Madame x is Francophone. Thus, obtaining answer n indirectly implies that Madame x has $\sim85.7\,\%$ chances to be Francophone. With the larger scale, on the right,
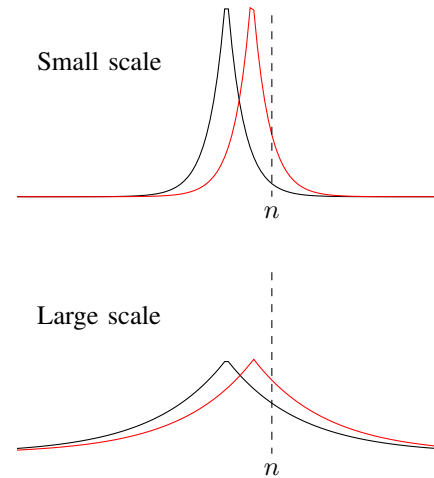

Fig. 6.   Different magnitudes of noise. *The black and red curves in each picture represent the probability of answer if Madame x is non-Francophone or Francophone, respectively. The vertical bar represents a possible (and arbitrary) answer n.*

answer n is about 1.3 more likely to occur if Madame x is Francophone, inducing "only" a belief of $\sim56.5\,\%$ that she is Francophone. A convenient property of Laplacian noises is that the ratio between both curves does not depend on the exact position of n. If you choose n further right, zoom in and measure again the ratio, it remains the same. The left side of the curves behaves symmetrically; as for the middle section between both peaks, it exhibits a varying ratio, but its value necessarily remains smaller than those on the right and left sides.

It is therefore possible to *bound* the amount of leakage such a query induces by choosing a desired noise magnitude. Now, one should keep in mind that while larger magnitudes of noise mean safer data, they also mean less accurate answers to the researchers. There is actually a tripartite tradeoff between *noise*, *utility*, and the *number of queries* that the system can hold.

B. Detailed principle

Private data analysis research is at the confluence of cryptology, statistics, and database systems. It focuses on statistical exploitation of privately held data by means of dedicated mechanisms. Adding noise to a query is one of these mechanisms, referred to as *output perturbation*. This topic has seen a resurgence of interest in the past five years, mainly due to a conceptual shift in the definition of privacy.

Differential privacy: Access to a statistical database should not enable one to learn anything about an individual, given that its data is in [the tested sample], that could not be learnt otherwise.

This definition should be understood as follows. Consider that we know that 1) Madame x underwent an angioplasty, and 2) 60% of angioplasty patients are Francophones (from a count query at SC). The subsequent belief that Madame x has 60% chances of being a Francophone should *not* be considered as a leakage over Madame x's language, because the same belief could have been inferred with or without her being in the tested sample. (Such information is precisely what OLMC researchers attempt to learn.)

Based on this definition, a new line of research was developed [3, 4, 5, 6, 7] around the question of how to limit the amount of information one can learn – *leakage* – about a specific individual that cannot be learned by means of an encompassing sample. The neighbor-sample setting discussed previously represents a worst-case scenario in this regard because the difference between both outputs can be attributed to a single individual; the leakage generated in this setting is therefore considered as an *upper bound* on the risk the database is more generally exposed to.

1) Relation between noise and leakage: Let us first imagine that *only one* query is performed over the whole lifetime of the system. (The extension to multiple queries is rather straightforward and discussed in a second step.) The main result from the field of differential privacy is to teach us how to generate the noise so as to observe a desired bound on the leakage, assuming this bound is expressed using the 'neighbor-sample' setting. Concretely, once this bound (let us note it $param$) is decided upon, differential privacy tells us how to generate the noise so as to ensure that

$$\frac{P[answer(s_1) = n]}{P[answer(s_2) = n]} \leq param \qquad (1)$$

for any neighbor samples s_1, s_2 and any possible answer n. The scale of the noise is then determined based on $param$ and on the *sensitivity* of the considered query; in our case, a *count* query. (The differential privacy framework is very general and applies to other types of queries; we will discuss some of them later in this document.)

a) Sensitivity: The sensitivity of a query, noted Δq, is defined as the maximal difference in output that two *neighbor* samples can induce. Two neighbor samples in our case cannot induce a difference larger than 1; thus, the sensitivity of count queries is 1. Note that the sensitivity does not depend on the size of the sample. This may appear counter-intuitive, but is actually consistent with generating the noise irrespective of the sample size.

b) The formula: The main result from [7], subsequently simplified in [3], tells us that differential privacy

is guaranteed with a given leakage bound '$param$' if the noise added to each answer is generated according to a Laplace distribution of scale

$$b = (\Delta q)/log(param) \qquad (2)$$

This is usually written $Noise \sim Lap(0, b)$, where the first parameter simply indicates that the noise will be *centered* on the true result.

c) Number of queries: This formula composes very well in the case of a sequence of queries. Indeed, the above papers tell us that, in order to maintain the same level of leakage for a sequence of queries, it suffices to generate the noise as if that sequence was a *single* query whose sensitivity is the count of all the sensitivities. When all queries are of a same type, this comes to considering the scale

$$b = (\#q\Delta q)/log(param), \qquad (3)$$

where $\#q$ is the number of queries in the sequence and Δq is the sensitivity of every individual query. A convenient consequence is that the database holder can think of choosing '$param$' irrespective of the number of queries.

2) Choosing the leakage parameter: A minimum of leakage is necessary for utility. If for instance $param = 1$, which implies that the probability of answer must be exactly the same for any neighbor samples s_1 and s_2, the problem that arises is that any *non*-neighbor pair of samples can be indirectly connected through a *chain* of neighbor samples. By transitivity, choosing $param = 1$ therefore implies that the sample has no impact on the answer, which is nonsense. How to choose $param$ is context-specific, and generally not discussed in the above papers. We explore this question in the context of language queries at Statistics Canada.

a) Belief: From the point of view of SC, the easiest way to formulate an 'acceptable' level of leakage is to specify the maximal belief allowed on the language of an individual. Say, for example, that this limit is 80% (*i.e.,* we should never believe that Madame x has more than 80% chances to be Francophone or Anglophone).

b) Worst case: We are interested in determining what value of $param$ should be chosen under the worst possible assumptions (mostly unrealistic). These assumptions are:

1) The adversary has no interest in health research; its only purpose is to learn Madame x's language.
2) The adversary has sufficient access in order to waste *all* the 'capital' of queries offered by SC in that sole purpose.
3) The adversary is able to build samples in which the language of each individual but Madame x is known.

4) SC applies no filter on the queries.

Consider the same attack scenario as discussed in Section III-A, based on two neighbor samples s_1 and s_2 (the second being equal to the first, plus Madame x). Because of point 3, the adversary already knows the *exact* answer for s_1, and can therefore use all the credit of queries to play and replay s_2. Assume it eventually gets a final answer n. Whether n corresponds to an average, a count, or any other combination of all the answers does not matter: differential privacy tells us that this answer cannot be $param$ times more likely for s_2 than for s_1, which is the only thing that matters here.

Three cases are considered, depending on the position of n with respect to the two probability peaks: if n is on the right of the right peak, then Madame x is $param$ more likely to be Francophone than non-Francophone; and symmetrically, if n is on the left of the left peak, she is $param$ more likely to be Anglophone. In the middle section, she could be up to $param$ more likely – but not more, as already discussed – to be either. Let us assume that n is on the right of the right peak, while keeping in mind that the following reasonning applies symmetrically. This implies that

$$\frac{P[x \text{ is Francophone}]}{P[x \text{ is not Francophone}]} \leq param$$

$$\implies \frac{P[x \text{ is Francophone}]}{1 - P[x \text{ is Francophone}]} \leq param$$

$$\dots$$

$$\implies P[x \text{ is Francophone}] \leq param/(1 + param) \tag{4}$$

This formula can be used to decide what value of $param$ corresponds to the maximal belief allowed by policy; in the case of our $80\,\%$-example above, this corresponds to $param = 4$. We now combine Equations 3 and 4 to represent the three-dimensional tradeoff between maximal belief, noise scale, and number of queries allowed. A meaningful cut of this volume is shown on Figure 7(a).

The right picture (Figure 7(b)) represents several examples of noise scales. Essentially, it gives the intuition that the precision of an answer deteriorates quickly with the scale parameter. How this affects the quality of the results for a researcher depends on how large the samples are. If samples are large (say, larger than 10000), then a scale of 30, or even 50, may not be a problem. If, on the other hand, the samples sizes are in the order of 1000, then any scale larger than 20 or 30 will induce meaningless answers. (This also depends on the specific question investigated and how deep it involves the language determinant.)
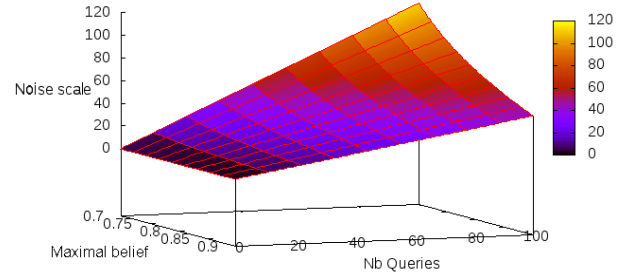
(a) Tradeoff between belief, noise scale, and number of queries. *The belief is over* 1 *(multiply by* 100 *for percentage).*
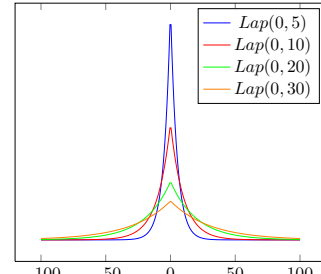
(b) Examples of noise scales.

Fig. 7.   Choice of parameters.

To close this illustrative discussion, let us consider that researchers require a scale no larger than 30, and SC tolerates a maximal belief of $80\,\%$, the tradeoff on Figure 7(a) tells us that up to $\sim$50 queries are still possible. This is not a lot, but keep in mind that this tradeoff represents a totally unrealistic worst-case scenario, where a single adversary can (and decides to) play *all* the credit of queries on a *single* sample, in which the language of *all* individuals but Madame x's is already known.

C. Discussion

What could be the consequences of relaxing some of the worst-case assumptions listed above? A first observation is that in our case, the researchers do not generate the samples directly; they do it through health criteria that are submitted to ICES (e.g. Figure 3), and the content of these samples is not known to the researchers. This implies that Assumption 3 is mostly unrealistic, and a given noisy answer will actually correspond to possibly more than two exact results, as illustrated on Figure 8. Further mathematical investigations are required to understand these implications precisely, but we can reasonably think that it would drastically lower the leakage that a query can possibly generate.

Another important aspect is to determine whether SC considers the couple ICES/researchers to be itself the adversary, or to be the potential victim of an external adversary. Put differently, does SC fear a direct misuse of the system, or a leak in data resulting from a normal-use? This point is crucial because the samples played in
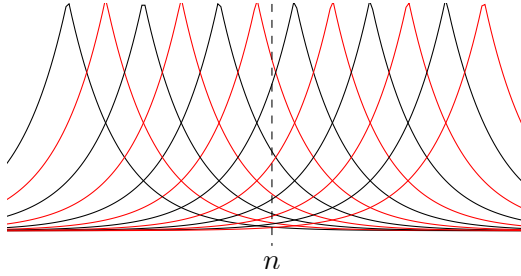
Fig. 8.   Guessing the exact answer from a noisy answer.

the context of a normal use are likely to generate much less leakage (intuitively, to the extent of a different order of magnitude).

As a third observation, Statistics Canada could easily filtrate the queries (Assumption 4) to prevent large intersection between the samples, and thereby drastically reduce the potential leakage of query combinations. Again, further mathematical investigation could be appropriate to explore, e.g., the correspondance between the size of intersection and the leakage. We believe that although difficult in the general case, this type of characterization remains reasonably feasible in the particular case of language-based count queries. We will contribute towards this direction if the option is concretely considered by the involved agencies.

In regards to all these considerations, it is not senseless to believe that the number of possible count queries could actually jump to several thousands, while keeping the leakage below any reasonable level, *e.g.,* much lower than a belief of 80%. If unfortunately these somewhat optimistic observations are not considered by the involved players (desiring to stick to the worst-case model), then there will still be a way to satisfy all the parts. The solution would consist in deploying a *non*-bounded service, in which the cumulated leakage is progressively computed as the queries are performed, and the service is shutdown when the leakage has reached a given threshold.

IV. RELAXING THE THREE INITIAL ASSUMPTIONS

A. Language

We assumed heretofore that the language of an individual could be described by a boolean value $\{Franco, Anglo\}$. Language is actually more subtle and requires looking at several aspects. In particular, the census data at SC includes the following variables: Mother tongue (for 100% of the population); Knowledge of official languages, Knowledge of non-official languages, First official language spoken, Language spoken at home, and Language of work (for 20% of the population). Also, some of these variables may include other options than

French and English, as well as a possible combination of several languages. It is clear that a mere *count query* cannot precisely render the ratio of "Francophones" in a given sample; at the most it could for example count the number of persons who report French as mother tongue and *not* English (the census form allows to report several mother tongues). Fortunately, the differential privacy framework can easily be applied to more complex queries, and in particular, *histograms*. Histogram queries typically return a set of counts, each corresponding to a disjoint property. They consequently seem to be an appropriate tool to perform less naive language queries. As explained in [3], the sensitivity of histograms is 2 because two neighbor samples cannot differ in more than two counts (and by more than 1 in each). Given a desired leakage parameter, and noise scale, the number of histogram queries that a system can hold is thus half that for count queries. Note that, here, the noise must be added to *each* count in the histogram, which has to be taken into consideration.

B. Linkage technicalities

So far, we have assumed that a correspondence was *technically* feasible between an individual and its language variables at Statistics Canada. In fact, the unit of census is not the individual, but the *household*. Both the short-form and the long-form census comprise a nominative field, but this field is apparently optional and is known to contain, occasionally, some exotic answers like 'Mickey Mouse'. Three options are possible here: 1) Assume that the persons living in a same household share similar language profiles, and then build the query samples based on postal addresses; 2) Rely on the cleaning operation that is currently performed on the census data at Statistics Canada, and which recently allowed to reach the level of 15% of records linkable nominatively (all from the long-form subset); or 3) Set up a complementary linkage, housed at SC, between the census data and the minimal amount of administrative data allowing to associate insurance numbers with the corresponding census records. (Note that this represents only a subset of what was done for Manitoba as part of a larger linkage including health data directly at SC [1].)

C. Generation of the samples at ICES

Health information is highly-sensitive, and ICES has a strong policy of confidentiality in this regard (see [11]). Concretely, the database held at ICES is sanitized by replacing all nominative information by anonymous identifiers. Whether these identifiers can technically and lawfully be reversed to the original information is a

question we are starting to investigate. It is likely that ICES maintains somewhere a private table with such associations. Thus, we believe that it is at least *technically* feasible. Such an operation should however be considered with utmost care. The feasibility also depends on what level of trust ICES grants to SC, even though SC does not need to know what health data is associated with a query (this is actually one of the advantages of the proposed mechanism). If need be, the body of research on cryptography techniques can be explored and leveraged. For instance, recent encryption methods have been proposed for the specific problem of querying databases using confidential inputs and criteria [12], that is in our case, preventing SC to learn what sample a given query is about. All these questions mainly fall within the area of ICES' Confidentiality Committee and Chief Privacy Officer, of which we hope to receive feedback on the present document. Somehow or other, we should be able to strive towards a technical solution that satisfies privacy constraints.

V. Concluding remarks

We discussed in this paper two possible mechanisms to enable the linkage of provincial health data with federal census data. Both mechanisms rely on the use of count queries, a type of operation that offers both powerful functionalities and assessable statistical risk. In the framework of our concern – health factors in official language minority communities – these mechanisms enable to supplement Ontario health data with linguistic variables from the 2006 Canadian Census. However, we believe this approach is more general and could certainly apply to different census variables than language (and different provinces than Ontario). Much remains to be done in this direction, but we hope this paper demonstrated at least that technical solutions do exist, and deserve to be explored.

VI. Acknowledgments

References

[1] C. Houle, J.M. Berthelot, P. David, M.C. Wolfson, C. Mustard, and L. Roos. Matching census database and Manitoba health care files. In *Intl. Workshop on Record linkage techniques*, page 305. National Academies, 1997.

[2] Statistics Canada Website. Policy on record linkage. http://www.statcan.gc.ca/record-enregistrement/policy4-1-politique4-1-eng.htm, 2011.

[3] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. *Theory of Cryptography (TCC'06)*, pages 265–284, 2006.

[4] K. Nissim, S. Raskhodnikova, and A. Smith. Smooth sensitivity and sampling in private data analysis. In *ACM Symposium on Theory of Computing (STOC'07)*, pages 75–84. ACM, 2007.

[5] C. Dwork. Ask a better question, get a better answer – a new approach to private data analysis. *Intl. Conference on Database Theory (ICDT'07)*, pages 18–27, 2006.

[6] C. Dwork and K. Nissim. Privacy-preserving datamining on vertically partitioned databases. In *Advances in Cryptology (CRYPTO'04)*, pages 134–138. Springer, 2004.

[7] A. Blum, C. Dwork, F. McSherry, and K. Nissim. Practical privacy: the SuLQ framework. In *ACM Symposium on Principles of Database Systems (PODS'05)*, pages 128–138. ACM, 2005.

[8] M. Hardt and K. Talwar. On the geometry of differential privacy. In *Proceedings of the 42nd ACM symposium on Theory of computing*, pages 705–714. ACM, 2010.

[9] A. Smith. Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the 43rd annual ACM symposium on Theory of computing*, pages 813–822. ACM, 2011.

[10] D. Kifer and A. Machanavajjhala. No free lunch in data privacy. In *Proceedings of the 2011 international conference on Management of data*, pages 193–204. ACM, 2011.

[11] Privacy at ICES. http://ices.queensu.ca/privacy.html, 2011.

[12] S.J. Stolfo and G. Tsudik, editors. *IEEE Security & Privacy Magazine, Special Issue on Privacy-Preserving Sharing of Sensitive Information*. IEEE Computer Society, July-Aug 2010.