Towards Practical Privacy-Preserving Life Cycle Assessment Computations

Cetin Sahin, Brandon Kuczenski, Omer Egecioglu, Amr El Abbadi
University of California, Santa Barbara
{cetin, omer, amr}@cs.ucsb.edu, bkuczenski@bren.ucsb.edu

ABSTRACT

Life Cycle Assessment(LCA) is crucial for evaluating the ecological sustainability of a product or service, and the accurate evaluation of sustainability requires detailed and transparent information about industrial activities. However, such information is usually considered confidential and withheld from the public. In this paper, we present a study of privacy in the context of LCA. The main goal is to explore the privacy challenges in sustainability assessment considering the protection of trade secrets while increasing transparency of industrial activities. To overcome privacy concerns, we apply differential privacy to LCA computations considering the idiosyncratic features of LCA data. Our assessments on a specific real-life example show that it is possible to achieve privacy-preserving LCA computations without losing the utility of data completely.

1. INTRODUCTION

One of the greatest challenges facing society is to ensure that the industrial goods and services required by a growing population can be met sustainably and equitably. Industrial Ecology (IE) is the study of resource requirements and the social and ecological implications of industrial activities. One primary technique in IE is life cycle assessment (LCA), a standardized methodology for estimating the environmental implications of products or services [3].

Preparing an LCA requires access to a database of information about the inventory requirements and environmental emissions of industrial processes, called a life cycle inventory (LCI) database. Preparing an accurate and comprehensive LCI database is a tremendous task and the development and maintenance of these resources is an ongoing challenge. Because industrial processes are typically undertaken in a competitive economic context, the operators of these processes would like to prevent potential competitors from learning sensitive information about their activities. Inventory data about industrial processes is usually considered to be confidential, and therefore is often not available freely. This type of information is nonetheless required to accurately assess environmental impact. Hence, the historical development of LCA has long been intimately bound to questions of confidentiality [4]. Despite its centrality to LCA, data privacy in the LCA domain has not been formally considered.

In this paper, we formulate the LCA computation in a way that introduces a privacy model, and consider possible threat models and attacks that could result in an adversary learning private data. Our goal is to provide the data security community with a real sense of the challenges faced by practitioners in the field of IE. We explore a particular problem in LCA and discuss the privacy issues and possible trade-offs between increase transparency by industrial companies and privacy protection of trade secrets that preserve competitive edge. The results of our attacks justify the concerns over publishing inventory data about industrial processes without any security provisions. To tackle this problem, we apply privacy techniques to LCA computations and illustrate their usage on a specific real life example. Our evaluations over a real life example highlight that it is possible to achieve privacy-preserving LCA publication without losing too much utility on the published data while ensuring privacy using differential privacy.

2. THE LCA AGGREGATION PROBLEM

LCA describes the delivery of a product or service as a network of industrial *unit processes*. Each unit process represents one form of industrial activity. Each edge in the network indicates a *flow* from one process to another, or between one process and the environment. Flows between a process and the environment are called *elementary* flows and may have environmental impacts. LCA studies distinguish between a *foreground model*, which represents the activities under scrutiny, and a *background model*, which represents the operations of the broader economy [8]. Private data are typically contained in the foreground model.

An *LCA aggregation study* can be described as three sequential matrix multiplications with respect to a background database B_x[5]. B_x is an $m \times n$ matrix that maps a set of n background processes to a set of m elementary flows. The foreground model is made up of a set of p foreground processes, each of which is defined by its dependencies on the n background processes. These are described in an $n \times p$ dependency matrix A_d, which comprises the study's private input data. Here w is a p-element weighting vector that specifies the relative significance of the different foreground processes. The first multiplication aggregates the foreground model into a weighted dependency vector a_p:

CODASPY'17 March 22-24, 2017, Scottsdale, AZ, USA

ACM ISBN 978-1-4503-4523-1/17/03.
DOI: http://dx.doi.org/10.1145/3029806.3029845

$$a_p = A_d \cdot w \tag{1}$$

The dependency vector a_p is then applied to the background database to determine an emission vector b:

$$b = B_x \cdot a_p \tag{2}$$

The vector b reports the aggregate amounts of different emissions released into the environment. The results of this computation are characterized with respect to a set of t environmental impact categories, represented by multiplication with a $t \times m$ characterization matrix E.

$$s = E \cdot b \tag{3}$$

This multiplication results in a set of t impact scores s, which are the final results of the study.

The current practice in the IE community is to make the result of the study s publicly available, so that their product system can be compared to other competitive product systems. However, it is difficult to evaluate the significance of the elements of s without knowing something about b. For instance, an independent researcher making a critical evaluation of s may wish to know whether a given environmental emission's contribution in b was significant. Additionally, some research questions may require a practitioner to supply their own E matrix, which is not possible if b is not disclosed.

3. CONFIDENTIALITY & PRIVACY ISSUES

To verify the validity of practitioners' concerns about publishing b, we here investigate the possible information leakage from the publication of b. In other words, how much of a_p can be recovered when b is published, given that B_x is public?

3.1 Industrial Ecology Privacy Concerns

The operations of an LCA aggregation study are sequential matrix multiplications. If B_x is a nonsingular (invertible) matrix, there exists a unique inverse denoted by B_x^{-1}, i. e., $B_x \cdot B_x^{-1} = B_x^{-1} \cdot B_x = I$. Then, Equation 2 has a unique solution, $a_p = B_x^{-1} \cdot b$. This might be seen as a justification for the concern not to publish b along with impact scores, s. However, B_x in LCA is a singular matrix most of the time, which means it is not invertible and a_p cannot be solved directly from Equation 2. Is this enough to ensure privacy guarantees?

The answer to this question is unclear. The concept of Moore-Penrose pseudoinverse of matrices [6], generalizes the notion of a nonsingular (invertible) matrix and makes it applicable to singular matrices. This concept is useful when searching for an optimal approximation of a set of linear equation solutions like $A \cdot x = y$, where A is a known $m \times n$ matrix, y is a column vector with m components and x is an unknown column vector. This approach can be directly applied in the LCA study to reveal the secret a_p vector with some approximation.

3.2 Revealing Industry Secrets

The Moore-Penrose pseudoinverse [6] guarantees a unique solution to x when A has a full column rank. In the context of LCA, to the best of our knowledge, having a full column rank in B_x is very rare, which leads to an infinite number of solutions for the linear system. One can claim that having an infinite number of solutions for x will create enough ambiguity and an adversary will not be able to distinguish which x is close to the original one. However, our empirical studies over a real LCA study disprove this and show that one can solve the linear system approximately close enough using the Moore-Penrose pseudoinverse. Therefore, we need to ensure the security of publication to prevent an adversary from recovering the solution even with the usage of Moore-Penrose inverse. In the context of privacy-preserving data publication, differential privacy is a canonical technique due to its strong privacy guarantees and capability to release useful aggregation information. Given that an LCA study is an aggregation problem, we propose differentially private LCA publications.

4. ACHIEVING LCA PRIVACY

Differential privacy provides a strong notion of privacy and is commonly used for statistical data publication [2]. It ensures that the removal or addition of a single record does not significantly affect the outcome of any analysis. Differential privacy can be achieved by the addition of random noise. The magnitude of the noise is chosen based on the sensitivity of a query function which considers the largest change in the output of the function with a change of a single record. Dwork [2] suggests using the Laplace mechanism to add noise to achieve differential privacy.

4.1 Differential Privacy for LCA Computation

Our goal is to perform differentially private LCA matrix multiplication in the form of Equation 2, where no adversary is able to recover a_p from the published b vector. Recall that B_x is a publicly known matrix. Each element in a_p represents a background process that is included in the production. The privacy goal is to make the publication such that either inclusion or exclusion of a specific background process has a negligible effect on the output, i.e., vector b. To achieve this goal, differential privacy might be applied by either perturbing the input or the output.

4.1.1 Input Perturbation

The initial approach to achieve differential privacy is to add noise to the input data itself. The straightforward approach is to generate a differentially private version of a_p, and then perform matrix factorization. In this case, the sensitivity of the publication considers the maximum change in all possible neighboring vectors.

Definition 1. Let $\mathbb{R}$ denote the set of real numbers. For $x_1,\ x_2 \in \mathbb{R}^d$, the sensitivity of the publication is:

$$\Delta f_1 = \max \parallel x_1 - x_2 \parallel_1 \tag{4}$$

for all x_1, x_2 differing in at most one element in the vector.

Now, we can formally define our differentially private vector publication mechanism.

Proposition 1. *The randomized mechanism M_K that outputs the following vector is ϵ-differentially private:*

$$M_K(x) = x + k \tag{5}$$

where k is a vector consisting of n independent samples drawn from the Laplace distribution function with a scale $\Delta f_1/\epsilon$, i. e., $Lap(\Delta f_1/\epsilon)$.

Recall that our motivation is to publish b, not a_p. Using M_K, it is possible to publish an ϵ-differentially private a_p. Now, this version of a_p can be used to compute b.

Proposition 2. *Given a public $A \in \mathbb{R}^{m\times n}$ and private $x \in \mathbb{R}^n$, the randomized mechanism M_{F_1} that performs the following operation ensures ϵ-differentially privacy for x:*

$$M_{F_1}(A, x) = A \cdot M_K(x) \tag{6}$$

4.1.2 Output Perturbation

To achieve differential privacy by perturbing the output, the desired differentially private mechanism initially computes the function, and then adds noise to each element of the computed output to obtain differentially private publication.

Definition 2. Let $\mathbb{R}$ denote the set of real numbers where $A \in \mathbb{R}^{m\times n}$ and $x \in \mathbb{R}^n$. A matrix multiplication function f:$\mathbb{R}^{m\times n} \times \mathbb{R}^n \to \mathbb{R}^m$ is defined by:

$$f(A, x) = A \cdot x \tag{7}$$

Definition 3. For $x_1, x_2 \in \mathbb{R}^n$, $A_1, A_2 \in \mathbb{R}^{m\times n}$, the sensitivity of $f(A, x)$ is:

$$\Delta f_2 = \max \parallel f(A_1, x_1) - f(A_2, x_2) \parallel_1 \tag{8}$$

for all x_1, x_2 differing in at most one element.

In the proposition below, we define a differentially private matrix multiplication mechanism.

Proposition 3. *Given a matrix multiplication function $f(A, x)$, the randomized mechanism M_{F_2} that outputs the following vector is ϵ-differentially private:*

$$M_{F_2}(A, x) = f(A, x) + k \tag{9}$$

where k is a vector consisting of m independent samples drawn from the Laplace distribution function with a scale $\Delta f_2/\epsilon$, i. e., Lap$(\Delta f_2/\epsilon)$.

Data dependent sensitivity: Although having a data independent sensitivity computation is a desired feature in differentially private publications, the sensitivity computations in our context are data dependent. In theory, the sensitivities, Δf_1 and Δf_2, are unbounded and can be infinity. Given this fact, differential privacy might be considered as an inappropriate methodology. However, this is not the case. LCA data has its own characteristics like sparsity, data distribution, which make differential privacy work in practice for the LCA computations. Due to lack of space, we skip the details. Please refer to the full version of the paper for detailed discussion and proofs [7].

5. EVALUATION OF PRIVACY-PRESERVING LCA COMPUTATION

We conducted experiments over a real LCA study for *distillers grain* using U.S. Life Cycle Inventory (USLCI) [1]. This study contains 39 background processes and 378 elementary flows. The distinctive property of this data set is the very broad range of numbers, i.e., from 10^{-15} to 10^3.

Attack against LCA publication: The attacker develops its attack by computing the Moore-Penrose pseudoinverse of B_x. The rank of B_x is 29 -not a full column rank-. This means the solution to the $B_x \cdot a_p = b$ linear system is not unique, and there is an approximate solution. It is reasonable to assume that an expert in the field has enough background knowledge to estimate which processes are included in the computation pretty well. With such background knowledge, the adversary can recover almost 82.05% of a_p. This outlines the power of the pseudoinverse approach in the context of LCA domain. Publishing b without any privacy technique has severe security issues, which justifies the concerns over making b public in the LCA community.

Differentially Private LCA Computation: We perform differentially private LCA computations using the mechanisms introduced in Section 4.1. When $\epsilon = 1$, M_{F_1} publishes a perturbed b, whose 165 elements out of 378 (44%) are approximately close within the threshold of 10^{-10}, i.e., the absolute difference between the original and computed entries is less than 10^{-10}. This is a good sign of utility. In addition, an attacker cannot approximate any element of a_p within a threshold of 10^{-10}. M_{F_2} does not provide as good utility as M_{F_1} regarding individual emission analysis. However, it delivers better utility if the analysis contains aggregate computations. Similar to the case with M_{F_1}, the attacker is not able to approximate any entries in a_p. Our study explores a normalization technique to decrease the utility loss. The gathered results justify the effectiveness of such an optimization by providing better utility for both M_{F_1} and M_{F_2} without sacrificing privacy (refer to the full version of the paper for detailed results [7]).

6. CONCLUSION

In this paper, we presented a study to explore the privacy concerns over publicizing industrial activities in the form of LCA computations. Our empirical studies show that the application of privacy-preserving techniques is required to preserve the privacy of private data and differentially private LCA computations ensure strong privacy while revealing useful information for analysts.

ACKNOWLEDGMENT

This work is funded by an NSF grant CCF-1442966.

7. REFERENCES

[1] U.S. Life Cycle Inventory Database, 2012. National Renewable Energy Laboratory, 2012. Accessed March 11, 2016: https://www.lcacommons.gov/nrel/search.

[2] C. Dwork. Differential privacy. In *ICALP 2006, Proceedings, Part II*, pages 1–12. Springer Berlin Heidelberg, 2006.

[3] G. Finnveden, M. Z. Hauschild, T. Ekvall, J. Guinée, R. Heijungs, S. Hellweg, A. Koehler, D. Pennington, and S. Suh. Recent developments in life cycle assessment. *Journal of Environmental Management*, 91(1):1–21, 2009.

[4] R. Frischknecht. Transparency in LCA-a heretical request? *Int J LCA*, 9(4):211–213, jul 2004.

[5] B. Kuczenski. Partial ordering of life cycle inventory databases. *The International Journal of Life Cycle Assessment*, 20(12):1673–1683, Oct 2015.

[6] E. H. Moore. On the reciprocal of the general algebraic matrix. *Bulletin of the American Mathematical Society*, 26:394–395, 1920.

[7] C. Sahin, B. Kuczenski, O. Egecioglu, and A. El Abbadi. Towards Practical Privacy-Preserving Life Cycle Assessment Computations. Technical report. January 2017. https://www.cs.ucsb.edu/research/tech-reports/2017-01.

[8] A.-M. Tillman. Significance of decision-making for LCA methodology. *Environ. Impact Assess. Rev.*, 20(1):113 – 123, 2000.