

Estimating Missing Unit Process Data in Life Cycle Assessment Using a Similarity-Based Approach

Ping Hou,^{†,‡} Jiarui Cai,[†] Shen Qu,[†] and Ming Xu^{*,†,§}

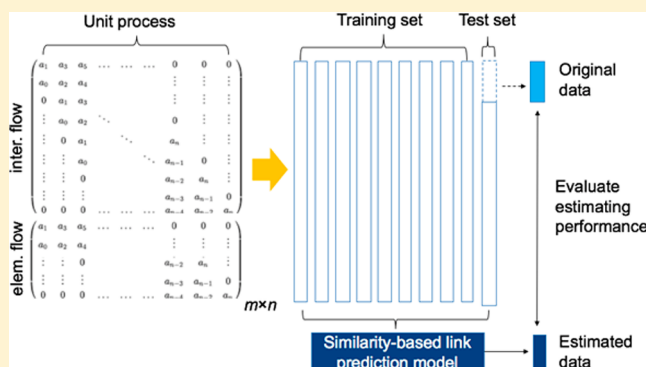
[†]School for Environment and Sustainability, University of Michigan, Ann Arbor, Michigan 48109, United States

[‡]Michigan Institute for Computational Discovery and Engineering, University of Michigan, Ann Arbor, Michigan 48104, United States

[§]Department of Civil and Environmental Engineering, University of Michigan, Ann Arbor, Michigan 48109, United States

Supporting Information

ABSTRACT: In life cycle assessment (LCA), collecting unit process data from the empirical sources (i.e., meter readings, operation logs/journals) is often costly and time-consuming. We propose a new computational approach to estimate missing unit process data solely relying on limited known data based on a similarity-based link prediction method. The intuition is that similar processes in a unit process network tend to have similar material/energy inputs and waste/emission outputs. We use the ecoinvent 3.1 unit process data sets to test our method in four steps: (1) dividing the data sets into a training set and a test set; (2) randomly removing certain numbers of data in the test set indicated as missing; (3) using similarity-weighted means of various numbers of most similar processes in the training set to estimate the missing data in the test set; and (4) comparing estimated data with the original values to determine the performance of the estimation. The results show that missing data can be accurately estimated when less than 5% data are missing in one process. The estimation performance decreases as the percentage of missing data increases. This study provides a new approach to compile unit process data and demonstrates a promising potential of using computational approaches for LCA data compilation.



INTRODUCTION

Life cycle assessment (LCA) measures the environmental impacts of a product in its whole life cycle, including resource extraction, raw materials processing, manufacturing, transport, use, and disposal.¹ By covering all stages of a product life cycle and a wide range of environmental impacts, LCA can help guide policy and technology development to avoid environmental burdens shifting among different life cycle stages. Increasingly, LCA has become an important tool in environmental policy and voluntary actions around the world, supporting decision-making toward sustainability.^{2,3}

An LCA model of a product is an ensemble of interconnected unit processes and system processes, respectively represented by foreground and background data. Foreground data quantify intermediate flows (i.e., materials/energy transmitted between unit processes) and elementary flows (i.e., resource from the environment and emission/waste released to the environment) associated with each unit process. Background data are aggregated life cycle inventory (LCI), normally provided by LCI database as system processes, which only contain elementary flows since all intermediate flows are traced back to resource extraction. On the basis of the product's LCA model, given a functional unit (e.g., produce 1 kg particular product), we can calculate the aggregated LCI of the product. The LCI results

multiplying with the corresponding characterization factors (i.e., the relative impact of LCI) are the characterized LCA results, which are the final results of an LCA study.

Collecting unit process data is fundamental to LCA study, but it is time-consuming, and needs large amounts of data, such as raw material inputs, energy use, the ratio of the main products to byproducts, production rates, and releases of emissions and waste. LCA practitioners often need to collect foreground and background data from a variety of sources, including direct reports from operations (e.g., meter readings, operation logs/journals), publications, government statistics, and LCA databases. In particular, LCA databases that provide LCA data for common processes are often the major sources for LCA data.⁴ Although convenient for LCA practitioners to use, LCA databases still largely rely on collecting empirical data from various sources, requiring a significant investment of human and capital resources.^{2,3}

In response to the difficulties facing collecting LCA data, there have been efforts to computationally estimate LCA data instead

Received: October 19, 2017

Revised: March 26, 2018

Accepted: March 30, 2018

Published: March 30, 2018

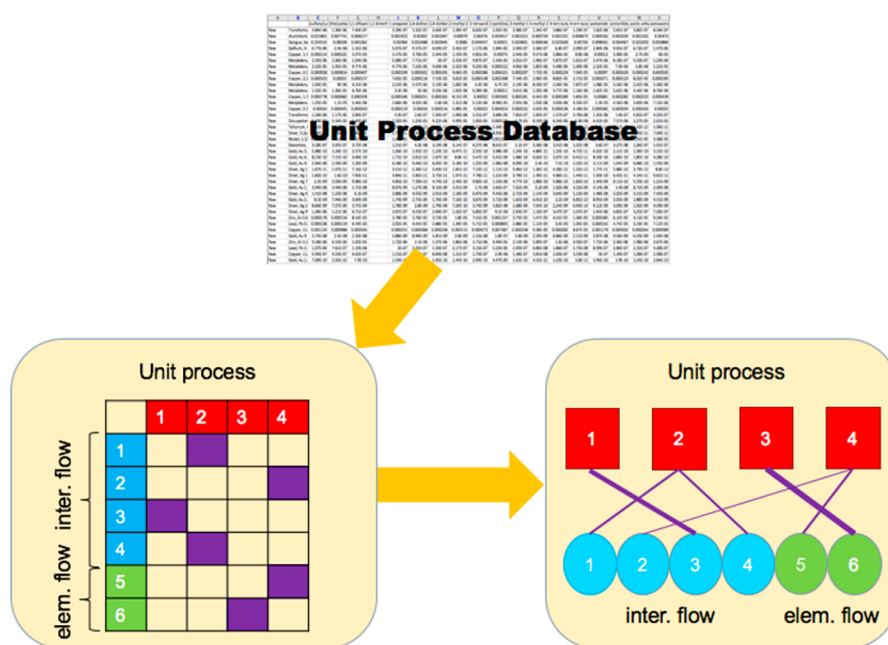


Figure 1. Data structure of a unit process database.

of relying on empirical data.^{5–11} For example, Wernet et al. developed molecular-structure-based models that use molecular features of chemicals as input to estimate chemicals' characterized LCA results using neural networks.^{5,6} Song et al. also used neural networks to estimate LCA results for chemicals, but with more layers in the neural networks.¹² In addition, given that electronic products are often designed as a bill of materials (BOM) to record the use of raw materials, components, and assemblies, some have proposed automating processes to assign impact factors to inventory components in the BOM to help the sustainable design of electronic products.^{7–9} Similarly, for buildings, studies have characterized the relationship between material uses and their environmental impacts to create new designs and optimize the energy use of buildings.^{10,11} However, all the previous studies still rely on collecting large amounts of information, such as chemical reaction equations, process characteristics of chemicals, and design details of electronic products and buildings. Therefore, these works only apply to specific products and rely on extensive domain knowledge. There is still a need to develop a convenient, computational approach to estimate LCA data with broader applications to a wide range of products. Suh and Huppes developed the Missing Inventory Estimation Tool (MIET) using extended input–output analysis. MIET is based on the national accounting system, thus includes the entire national economy; but the application is limited due to the coarse resolution of the input–output tables and product prices are required to convert monetary units to physical units.¹³ Other studies have mostly focused on estimating characterized LCA results, except Suh and Huppes¹³ estimate both LCA results and unit process data. In our study, we estimate unit process data which are the basis of all the LCA results.

In this study, we developed a similarity-based computational framework for estimating missing unit process data solely based on limited known data, without relying on additional empirical data. In particular, we used similarity-based link prediction. Link prediction is a branch of the emerging network science to predict missing links in a network based on limited observations.¹⁴ Link prediction has mostly been applied in recommendation

systems,^{15–18} such as in e-commerce sites to recommend products/services to likely customers.¹⁹ Link prediction has also been applied in analyzing social networks, such as characterizing the structure of literature citation networks,²⁰ predicting collaborations in coauthorship networks,¹⁴ and detecting relationships among terrorists.²¹ Viewing the unit process data as a network allows us to use link prediction to explore the interrelationship between unit processes. We used the ecoinvent 3.1 unit process data sets (UPR) to test this method. In particular, we first used our method to estimate a subset of the UPR data using the other subset as the training data. We then compared the estimated data with the original UPR data to evaluate the performance of the method. The detailed description of methods and materials is in the next section.

METHODS AND MATERIALS

The intuition of our method is based on the fact that unit process data essentially represent the interrelationship of unit processes (by intermediate flows) and the interrelationship between unit processes and the environment (by elementary flows). The ensemble of such interrelationship characterizes the structure of the underlying technology network (or unit process network). If sufficient, observed unit process data, although not complete, can potentially be used to extract structural features of the underlying technology network. Such structural features, in turn, can be used to predict the structure of the unknown area of the technology network, which is equivalent to estimating the unknown data in the unit process database. In this work, we used the similarity of unit processes in a unit process database to characterize the structure of the unit process network and developed the computational model to estimate missing unit process data.

Unit Process Data Structure. Unit process data are commonly represented as matrices when used in matrix-based LCA models. As illustrated in Figure 1, a unit process database is a matrix with columns representing unit processes (e.g., production of 1 kWh electricity) and rows representing either intermediate flows (i.e., inputs required by each unit process from other unit processes) or elementary flows (e.g., water

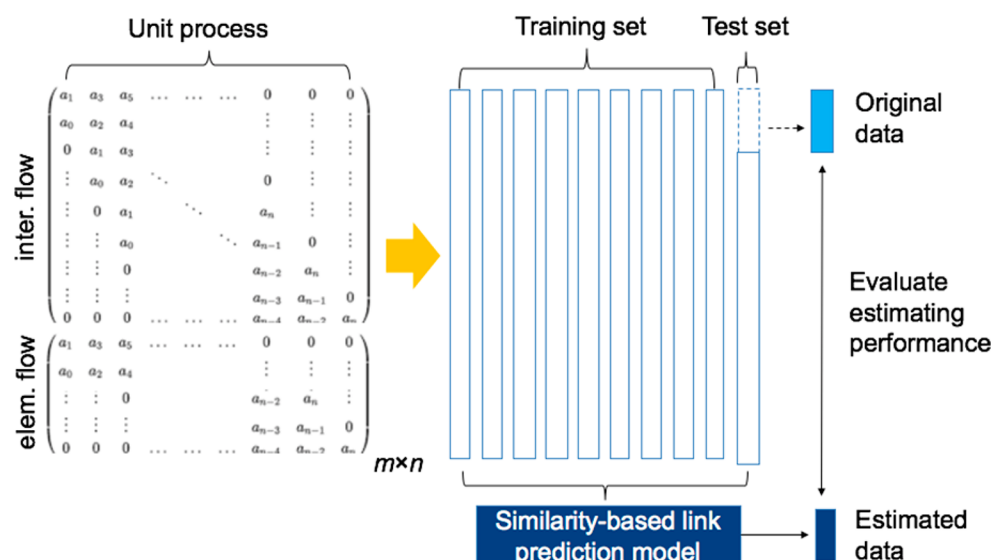


Figure 2. Methodological framework of similarity-based link prediction applied to estimating unit process data.

consumption, CO₂ emissions). Each element of the matrix indicates the amount of a particular type of intermediate or elementary flow (row) associated with the unitary output of a particular unit process (column), e.g., 1.07 kg CO₂ emissions per kWh electricity production in hard coal power plants. Therefore, this matrix is a combination of technology matrix (A matrix) and emission matrix (B matrix).²²

A unit process database can also be represented as a network, using the unit process matrix as the adjacency matrix (Figure 1). In particular, there are two types of nodes (or vertices) in a unit process network respectively representing unit processes and intermediate and elementary flows. Unit process nodes are connected with intermediate or elementary flow nodes by links (or edges) indicating how much and what type of flow with each unit process is associated. This network is a weighted, bipartite network,²³ as its links have strengths (the amount of flow) and nodes are divided into two disjoint sets. The network representation allows studying structural features of complex systems.^{24,25} Representing a unit process database as a network allows identifying the features of its structure which is then used to estimate missing unit process data.

Link Prediction and Similarity-Based Link Prediction.

Many link prediction applications use similarity-based methods.²⁶ As the name suggests, similarity-based methods first measure the similarity (or proximity) between each pair of nodes in the network. For bipartite networks, the similarity is measured for the same type of nodes. Two nodes that are similar tend to have similar patterns of linkages with other nodes in the network. On the basis of appropriate measures of similarity, we can then evaluate the likelihood of unknown links that exist for a node by comparing it with other similar nodes. We apply the principle of similarity-based link prediction in this work to estimate missing unit process data. Note that in unit process networks, only predicting the existence of links between processes is not enough. We also need to predict the strength of particular links. Although this is different from simply applying existing link prediction methods that are mostly developed for unweighted networks, the same principles still apply.^{27,28}

Methodological Framework. As shown in Figure 2, we use a reputable unit process database as the complete, observed data set, which is an $m \times n$ matrix including m types of intermediate

and elementary flows and n types of unit processes. For each process j (column $j \in [1, n]$ in the matrix), when we use it as the test set, the rest of the matrix becomes the training set. We then randomly select p ($1 \leq p < m$) number of data from the test set (column j) and assume they are missing. We use the training set to estimate those missing data based on the similarities of the remaining data in the process j with the k ($1 \leq k \leq n - 1$) most similar processes (details described below). Finally, we compare the estimated data with the original data to evaluate the performance of the method. To measure the effectiveness of our method, this procedure is repeated for each process being used as the test set, the same as in leave-one-out-cross-validation (LOOCV). The overall performance of the method with respect to the selected unit process database can then be evaluated by averaging the performance metrics obtained each time (for details see eq 4).

For each test set with certain numbers of missing data, we use the following three steps to complete the estimation and evaluate the performance of the estimation.

1. Similarity Calculation. We compute the similarities of test set (process j) which has missing data with other processes in the training set by comparing the remaining portion of process j and the corresponding portion of each process in the training set. Although many methods are available to measure the difference between two vectors, we choose the Minkowski distance based on comparison of these available methods (Supporting Information, SI Text S1 and Table S1). In particular, the Minkowski distance between the remaining portion of the test set and the corresponding portion of each process in the training set is calculated as

$$d_{ij} = \left(\sum_{t=1}^{m-p} |a_{ti} - a_{tj}|^q \right)^{1/q} \quad (1)$$

where t indexes intermediate and elementary flows, $m-p$ is the total number of flows minus the number of the missing data (i.e., number of known data in process j), a_{ti} is the flow t in the training process i , a_{tj} is the flow t in the test process j , and q is the parameter in the definition of Minkowski distance. We can get different measurements of the distance by adjusting q . During the training, we find the best q which achieves the lowest estimation

errors for each training data set. The similarity of the known portion of process j and the corresponding portion of process i , s_{ij} , is then calculated based on their distance d_{ij} .

$$s_{ij} = \frac{1}{d_{ij} + 1} \quad (2)$$

The larger the s_{ij} is, the more similar the two processes are. If one process in the training set is identical with the process in the test set, then their distance d_{ij} is 0 and their similarity s_{ij} becomes 1.

2. Missing Data Estimation. Each missing data point e_{ij} in the test set process j is estimated by averaging the corresponding data in the k most similar processes weighted by their similarities, which are calculated by eqs 1 and 2.

$$e_{ij} = \frac{\sum_{i=1}^k a_{ti} s_{ij}}{\sum_{i=1}^k s_{ij}} \quad (3)$$

where k ($1 \leq k \leq m - 1$) represents the number of most similar processes in the training set used to estimate the missing data in the test set and a_{ti} is the corresponding flow t of the i -th similar process when the training processes are ranked in descending order of similarity. For every set of missing data, there are $m - 1$ different estimations with k ranging between 1 and $m - 1$.

3. Performance Measurement. We evaluate the performance of the model by comparing the estimated data e_{ij} with the original data a_{ij} using mean percentage error (MPE):

$$\text{MPE} = \sqrt{\frac{\sum_1^p (a_{ij} - e_{ij})^2}{\sum_1^p a_{ij}^2}} \quad (4)$$

where p is the number of the missing data. Lower MPE indicates more accurate estimation of the missing data.

One important assumption of our method is that the observed unit process data we use to estimate the missing data should be complete. In other words, the applicability of our method largely depends on the completeness and quality of the observed unit process database. Our method does not intend to replace primary data collection for unit processes, but is a complementary approach when primary data are not available.

Data. We used ecoinvent 3.1 database²⁹ as a reputable database to test our method. The ecoinvent database is perhaps the most widely used LCA database, which comprises data for thousands of common unit processes. There are three models in ecoinvent including default model, cutoff model, and consequential model. The three models have the same matrix structure, except they use different methods to deal with coproducts and wastes. For each model, ecoinvent provides three data sets including unit process data sets (UPR), aggregated life cycle inventories (LCI), and calculated impact assessment results (LCIA). UPR records the data for energy/resource inputs and emission outputs of a process. Aggregated LCI converts all upstream UPR data into the life cycle inputs and outputs of a process. LCIA is the characterized LCA results by categorizing energy and resource uses and emissions into various categories of environmental impacts. In this study, we used the UPR data in the default model to represent the underlying technology network of the ecoinvent database, as the aggregated LCI and LCIA data are essentially calculated based on UPR for users' convenience.

Ecoinvent 3.1 UPR database is a 13 201 by 11 332 matrix, corresponding to 13 201 types of flows (including 11 332 types of intermediate flows and 1869 types of elementary flows) and

11 332 unit processes. Because UPR only includes on-site energy and resource use and emission data for each process, most entries in the UPR matrix are zeros. Thus, the ecoinvent UPR database is a sparse matrix, in which only 9.8% of entries are nonzero.

The ecoinvent database implements inheritance for geography, which means a local process can be created as a child of the global parent process. The child process inherits all flows from the parent unless otherwise specified to ensure consistency of processes for the same activity in different regions.³⁰ Some local processes are generated as an exact copy of the global process with uncertainty adjusted. We kept only the parent process by removing the child processes specific for different locations. Empty rows and columns in the UPR matrix are also removed. As a result, the processed UPR matrix has 7029 intermediate and elementary flows (row) and 2546 processes (column).

Data in the ecoinvent database have very different orders of magnitude due to the nature of economic and environmental flows and the choice of units. For example, CO₂ emissions to the air can be in the order of 10⁻⁵–10¹ kg for the unitary output of a process, while lead discharges to the air for the same unitary output of the same process can be only in the order of 10⁻¹⁵–10⁻¹⁸ kg. If one of the flows has a relatively high order of magnitude, the similarity of unit processes will be dominated by this particular flow. In addition, the choice of units of processes also affects the order of magnitude. For instance, the data set of 1 km passenger car transportation and the data set of 1 t*km freight train transportation have very different orders of magnitude since the later converted the data to per metric ton freight being transported. Normalization sometimes is needed to represent data in similar orders of magnitude. In this paper, we define a specific procedure of matrix normalization (Text S1) and compare three different strategies: (1) normalization based on the complete UPR matrix; (2) normalization based on the training set to avoid introducing future information in the test set; and (3) without normalization. Table S1 compares the estimation results of using these three strategies. Overall, estimation without normalization offers the best results. This is because normalization, while making the data more regular, can actually lose important information from the raw data set. Such information can be useful to improve the estimation accuracy.

RESULTS AND DISCUSSION

Similarity of Processes. We calculated the similarities of each pair of processes in the ecoinvent 3.1 UPR data set. The resulting similarity matrix is a symmetric, square matrix with both rows and columns representing unit processes and elements standing for the similarities (s_{ij}) between pairs of processes. Because the processed matrix has 2546 processes, the similarity matrix is a square matrix of 2546 by 2546. It shows the similarities of each process with the other 2545 processes in the matrix. Each cell's value s_{ij} is calculated by eqs 1 and 2, showing the similarity of process i with process j . The smaller the s_{ij} is, the less similar process i is to process j . Figure 3 shows the heat map of the similarity matrix to demonstrate the disparity of similarities of each pair of processes. The grid-like pattern indicates that the process pairs have significantly different levels of similarity, which provides valuable information to extract the underlying structure of the data set. The processes in Figure 3 are ordered by International Standard Industrial Classification of All Economic Activities (ISIC) in which processes in the same industry are next to each other. We observe irregular distribution of similarities in Figure 3. Specifically, there are no light squares around the diagonal, meaning processes from the same industry do not

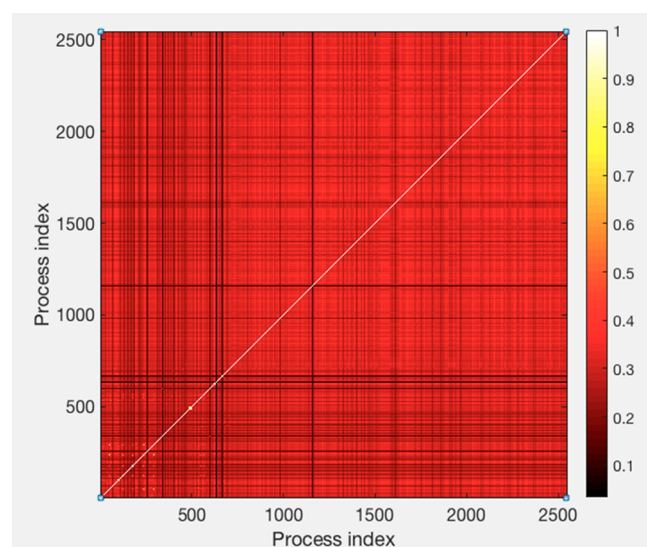


Figure 3. Heat map of the similarity matrix for the ecoinvent 3.1 UPR data set.

necessarily have similar intermediate and elementary flows. Therefore, using processes from the same industry to update missing intermediate and elementary flows as commonly done in LCI is in fact not always appropriate. Note that the similarities shown here are not the similarities we use to estimate the missing data. The similarities here are based on the complete data set. When we estimate missing data, similarities are calculated each time only based on the remaining data after the missing data are removed.

As an example, we find the ten most similar processes for the process “machine operation, diesel, < 18.64 kW, underground mining” (Table 1). The four most similar processes are all

Table 1. Ten Most Similar Processes for “Machine Operation, Diesel, < 18.64 kW, Underground Mining” (Index Number 2429 in Figure 3)

	process	similarity	index number in Figure 3
1	machine operation, diesel, < 18.64 kW, steady-state	0.9083	2428
2	machine operation, diesel, < 18.64 kW, low load factor	0.8925	2427
3	machine operation, diesel, < 18.64 kW, high load factor	0.8498	2426
4	machine operation, diesel, < 18.64 kW, underground mining	0.7843	2429
5	excavation, skid-steer loader	0.7098	294
6	excavation, hydraulic digger	0.7095	293
7	diesel, burned in building machine	0.7062	688
8	bale loading	0.7055	119
9	baling	0.7053	120
10	waste vapor barrier, flame-retarded	0.7052	1820

basically the same as the selected process, just with different conditions. The fifth to ninth most similar processes all involve the combustion of diesel as the selected process does. The tenth is a waste incineration process which also involves combustion.

Estimation Performance. We used mean percentage error (MPE) to evaluate the accuracy of estimating different numbers of missing data. Specifically, we tested missing 1%, 5%, 10%, and 20% of the total number of intermediate and elementary flows

(7029). Figure 4 shows the MPEs when different percentages of data are missing when k (the number of most similar processes) ranges from 1 to 2545 and q (the parameter in the distance function) ranges from 0.01 to 0.1. The MPEs are the average of the MPEs by making every process in the data set as testing data one by one. When fewer data are missing, the estimation MPEs are distributed in relatively narrow ranges (e.g., 1% and 5% data missing in Figure 4). This implies that most processes can be estimated relatively well except for a few outliers. When more data are missing (e.g., 10% data missing), the distribution of MPEs becomes much broader. This indicates the number of processes that are difficult to estimate becomes larger when more data become missing. However, when even more data are missing (e.g., 20% data missing), the MPEs are distributed again in a narrow range but with large values, which means when missing data exceeds a certain level, the missing data are generally hard to estimate. This is because, when more data are missing, the less information we can use to estimate those missing data and the similarity measures are getting less reliable in finding similar unit processes. MPE is the lowest when a few most similar processes are used for the estimation. However, when more processes are included, MPE values actually increase, because newly added processes are less similar and introduce more noises. Therefore, using more processes for the estimation does not necessarily mean lower MPE.

Figure 5 shows the distribution of the MPEs for all processes with respect to different percentages of data missing. On the basis of these distributions, we can choose the value of parameter q that has the best estimation performance. The parameter q in the Minkowski distance is essentially a general representation of distance function. When $q = 1$, it is Manhattan distance; when $q = 2$, it is Euclidean distance. Larger values of q place more emphasis on large differences in intermediate and elementary flows, because all differences are raised to the power of q . Consequently, the distance with higher q is strongly influenced by a single large difference in one flow. From Figure 5, we can see that smaller q generally tends to correspond to lower MPEs. Because we have in total 7029 flows, placing more emphasis on a small number of large differences is not helpful on the estimation.

Table 2 and Figure 6 show the distribution of the MPEs for all the processes with respect to the best q , which is 0.01. MPEs are distributed around the average value with relatively small standard deviations. These results show that the accuracy of estimation increases (i.e., average MPE decreases) when fewer data are missing. When 1% and 5% data are missing, we can estimate those missing data with a very high accuracy (i.e., the average MPE are almost near zero). Given that the data set includes many entries as zero, missing less than 5% data often means a few nonzero data points are missing. As a result, one or two processes in the training set that are most similar to the process in the test set can effectively dominate the estimation and make the estimation very close to the original values. An example of missing 1% data in a test set with high estimation accuracy is given in the SI (Text S3). When 10% data are missing, the average MPE becomes 39.32%. The average MPE exceeds 90% when more than 20% data are missing in one process. We did not test more scenarios of data missing since the result for 20% missing is already beyond acceptance and more data missing will only worsen the estimation results.

We also examined the computational resources required to complete the estimation on both single processor and multiple processes using parallel computing. As shown in the SI (Text S4

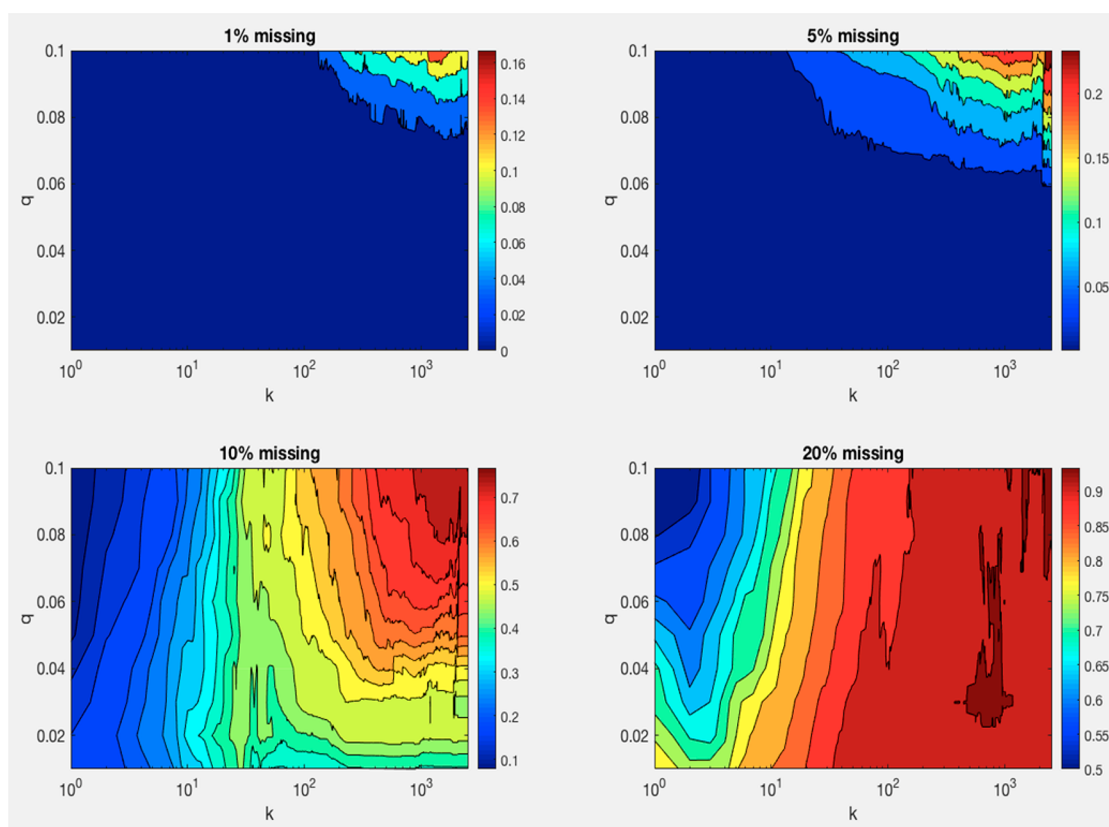


Figure 4. MPEs with respect to percentage of data missing, k (the number of most similar processes), and q (the parameter in the distance function).

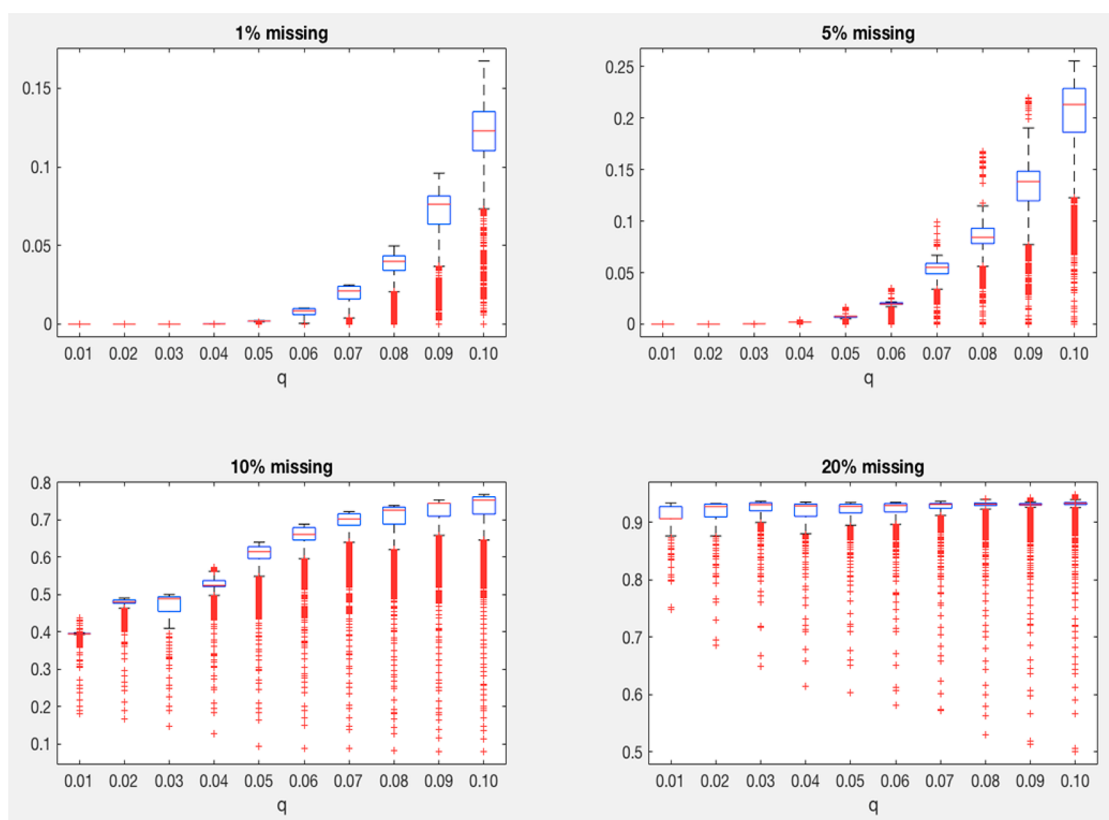
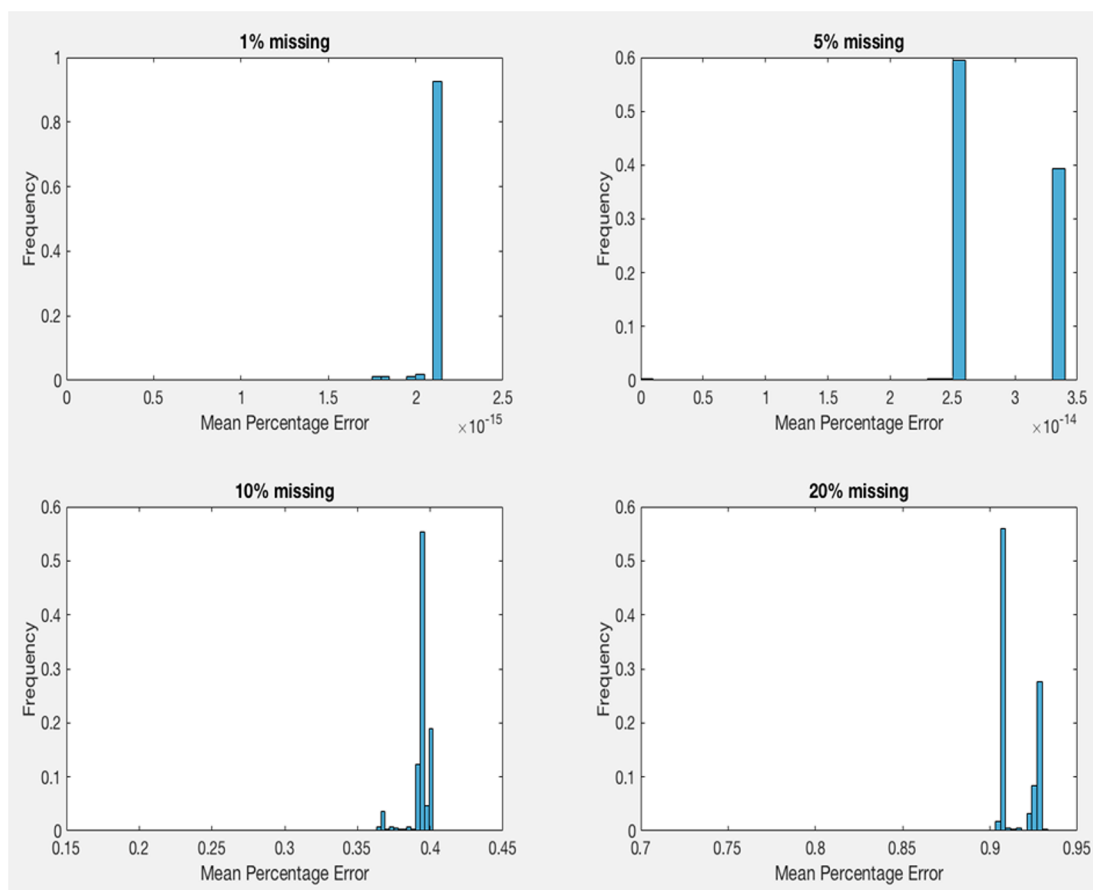


Figure 5. Distribution of MPEs with respect to the percentage of data missing and value of q (the parameter in the distance function).

Table 2. MPEs with Different Percentages of Data Missing

MPE	average	25% quantile	median	75% quantile	standard deviation
1% missing	$2.09 \times 10^{-13}\%$	$2.12 \times 10^{-13}\%$	$2.12 \times 10^{-13}\%$	$2.12 \times 10^{-13}\%$	$1.61 \times 10^{-14}\%$
5% missing	$2.85 \times 10^{-12}\%$	$2.54 \times 10^{-12}\%$	$2.54 \times 10^{-12}\%$	$3.35 \times 10^{-12}\%$	$4.39 \times 10^{-13}\%$
10% missing	39.32%	39.45%	39.46%	39.58%	1.26%
20% missing	91.39%	90.61%	90.61%	92.71%	1.37%

Figure 6. Histograms of MPEs when different percentages of data are missing with the best q .

and Table S3), overall the computational resource needed for implementing our method in ecoinvent 3.1 is manageable.

Theoretical Grounds. Although the computational results are promising, we further explored theoretical grounds that can explain the results. Essentially, the ecoinvent UPR data set is high-dimensional with 7029 variables (intermediate and elementary flows) but only 2546 samples (processes). Such high dimensional data often lie close to low dimensional subspaces. The intrinsic dimension of the matrix can be measured with the effective rank³¹

$$r(\Sigma) = \frac{\text{tr}(\Sigma)}{\|\Sigma\|} \quad (5)$$

where $\text{tr}(\Sigma)$ is the trace of a matrix, which is the sum of all the singular values of Σ , and $\|\Sigma\|$ denotes the largest singular value of Σ .

On the basis of eq 5, the effective rank of the processed data set is approximately 2. This means that the first two components explain 89.3% of the total variance. This is reasonable since some intermediate flows and elementary flows are highly related with each other. For example, processes with high energy consumption normally have high levels of greenhouse gas

emissions. This low rank characteristic of the ecoinvent 3.1 UPR data set allows us to estimate limited number of missing data with the underlying pattern of the data set and explains why our method works very well when missing 1%–5% data. We also recognize the diversity among the types of flows and the categories of unit processes. Such diversity could be the reason why the estimated results are not satisfactory when more data are missing.

Similarity Measurement. In our method, we solely use unit process data to measure the similarity between any pair of processes, while ecoinvent also categorizes unit processes based on their industrial classification. Intuitively, one would assume that processes in the same category would be more similar to each other. We calculate the similarity of each pair of processes and found the most similar process for each process. However, the results show that only approximately 32% unit processes are in the same category with their most similar process. For example, cement and clinker both belong to the construction industry. However, because clinker production is highly energy intensive but cement production is just to mix raw materials including clinker, clinker and cement production processes are very different despite being in the same industry. Our results

prove that their similarity is low (0.2308). Therefore, using processes in the same industry to update missing flows, as commonly done in LCA practice, is not always appropriate.

Case Study. To demonstrate the application of our method, we choose one process in the U.S. LCI database,³² “Diesel, combusted in industrial boiler”, to identify its possible missing flows. We calculate the similarities between this process and all the unit processes in ecoinvent. The most similar process in ecoinvent is “machine operation, diesel, < 18.64 kW, low load factor”. The descriptions of these two processes suggest strong similarity as well. The U.S. LCI process and its most similar process in ecoinvent have 17 flows in common. The MPE of estimating the 17 flows of the U.S. LCI process using its similarities with ecoinvent processes is 7.05%. Comparing the two processes, the ecoinvent process has 36 additional intermediate flows (e.g., lubricating oil) and elementary flows (e.g., ammonia and benzene emissions) (Text S5). This suggests the U.S. LCI process potentially misses data for these additional flows. Therefore, ecoinvent data can be used to estimate data for these additional flows for the U.S. LCI process.

Implications for LCA. We envision three major implications for LCA research and practice. First, empirical unit process data collection is expensive and time-consuming. With sufficient amount of known, trusted (to some extent) data, we show that the similarity-based link prediction approach can effectively estimate missing data with high levels of accuracy. The adoption of this approach will provide reasonably accurate unit process data when primary data are not available, with only a fraction of cost and time for collecting primary data.

Second, unit process data are collected from various sources with different quality. We can use a portion of the unit process data set that is trustworthy to estimate the less trustworthy data. By comparing the estimated results with the collected data, we can identify the data potentially with low quality. This result can help guide directions for improving data quality.³³ In ecoinvent, there are many missing intermediate and elementary flows that are represented as zeros. To investigate which of the current zero entries should in fact not be zeros, we apply our method on the zero entries when 1% data missing. We find that the estimated data for the zero entries are also zeros for 90% of the processes. The remaining 10% unit processes, for which estimated data are nonzeros, are generally nonexceptionally market processes, newly introduced in ecoinvent v3. The distinctive feature of these processes is that there is no transformation of materials happening, simply adding transport activities, wholesale and retail activities, and product losses in trade and transport. Missing data for market processes in ecoinvent are substituted by a simple market data set.³⁰ This is the reason our method estimated the zero entries as nonzeros. In other words, our method can be used to identify those missing or low-quality data and direct future data collection efforts.

Lastly, LCA databases are constantly expanding due to the addition of new unit processes from new technologies. It is often the case that the unit process data for new, emerging processes are incomplete. Our method can be used to estimate the incomplete data for a new process based on its similarities with other processes in a known LCA database.³⁴ Note that our method does not apply if there are no data at all for a new process, because we cannot compute the similarity and find the relationship between the new process and other processes. However, it is also very rare that we do not know anything about the new process except what it produces. At least, one should be able to know energy and key material uses for producing unitary

product from this process, which can potentially be used to estimate other intermediate and elementary flows.

Limitations. Although we envision broad applications of our method, it still has a long way to go to practical applications. As mentioned before, our method is based on the assumption that the observed unit process data used to estimate the missing data are complete. In practice, it is impossible to have an observed database that is complete. In fact, as the most comprehensive and widely used LCA database, ecoinvent still has many missing data that are simply filled with zeros.

Our results show that the estimation accuracy is lower when more data are missing (e.g., over 20%). This is because, when the number of known data is limited, the similarity calculated based on such limited data is no longer accurate and reliable. Therefore, our method is more applicable in the situation that most data are known and only a few data points are missing.

Despite all these limitations, we still believe our study provides a new direction for estimating unit process data for LCA. In future research, how to find the optimal choice of similarity measurements and the number of most similar processes used for estimation needs to be carefully examined.

Future Work. Our results and conclusions only apply to the ecoinvent 3.1 UPR data set. In addition, we need to test our method in other commonly used LCA databases, such as the Greenhouse Gases, Regulated Emissions, and Energy Use in Transportation (GREET) model³⁵ and the U.S. LCI Database.³² In particular, the ecoinvent database is one of the proprietary databases with comprehensive coverage; GREET has been developed for a particular sector (transportation); and the U.S. LCI Database is a national reference LCA database that provides industrial-representative LCA data for a particular country. Testing our method using these representative LCA databases will help understand the applicability of our method and its limitations.

In addition to the similarity-based link prediction method used in this study, other methods we have seen in the recent rapid development of data science can potentially also be used to estimate missing unit process data. Our future work will explore the potential applications of these methods in LCA data estimation.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.est.7b05366.

Comparison of the performance of the estimation using various methods to measure distances between processes; computational time needed for completing the estimation (PDF)

An example of estimating 1% missing data in a test process with high accuracy; a case study demonstrating the application of our method (XLSX)

■ AUTHOR INFORMATION

Corresponding Author

*E-mail: mingxu@umich.edu.

ORCID

Ming Xu: 0000-0002-7106-8390

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant 1554349. We thankecoinvent Centre for providing data and Gregor Wernet, Andreas Ciroth, Jarod Kelly, Sarah Boyd, Jon Dettling, Dali Wang, and Hongtao Wang for their guidance and feedback in developing this research. We also thank Advanced Research Computing, Technical Service (ARC-TS) at the University of Michigan for support on code optimization.

■ REFERENCES

- (1) Rebitzer, G.; Ekvall, T.; Frischknecht, R.; Hunkeler, D.; Norris, G.; Rydberg, T.; Schmidt, W. P.; Suh, S.; Weidema, B. P.; Pennington, D. W. Life cycle assessment Part 1: Framework, goal and scope definition, inventory analysis, and applications. *Environ. Int.* **2004**, *30* (5), 701–720.
- (2) Guinee, J. B. Life cycle assessment: past, present and future. In *International Symposium on Life Cycle Assessment and Construction: Civil Engineering and Buildings*; Nantes, France, Jul 10–12, 2012; pp 9–11.
- (3) Hellweg, S.; Milai Canals, L. Emerging approaches, challenges and opportunities in life cycle assessment. *Science* **2014**, *344* (6188), 1109–1113.
- (4) Frischknecht, R.; Jungbluth, N.; Althaus, H. J.; Doka, G.; Dones, R.; Heck, T.; Hellweg, S.; Hirschier, R.; Nemecek, T.; Rebitzer, G.; Spielmann, M. The ecoinvent database: Overview and methodological framework. *Int. J. Life Cycle Assess.* **2005**, *10* (1), 3–9.
- (5) Wernet, G.; Papadokostantakis, S.; Hellweg, S.; Hungerbühler, K. Bridging data gaps in environmental assessments: Modeling impacts of fine and basic chemical production. *Green Chem.* **2009**, *11* (11), 1826–1831.
- (6) Wernet, G.; Hellweg, S.; Fischer, U.; Papadokostantakis, S.; Hungerbühler, K. Molecular-structure-based models of chemical inventories using neural networks. *Environ. Sci. Technol.* **2008**, *42* (17), 6717–6722.
- (7) Sundaravaradan, N.; Marwah, M.; Shah, A.; Ramakrishnan, N.; Data Mining Approaches for Life Cycle Assessment. In *2011 IEEE International Symposium on Sustainable Systems and Technology (ISSST)*, 2011.
- (8) Ramakrishnan, N.; Marwah, M.; Shah, A.; Patnaik, D.; Hossain, M. S.; Sundaravaradan, N.; Patel, C. Data mining solutions for sustainability problems. *Ieee Potentials* **2012**, *31* (6), 28–34.
- (9) Marwah, M.; Shah, A.; Bash, C.; Patel, C.; Ramakrishnan, N. Using Data Mining to Help Design Sustainable Products. *Computer* **2011**, *44* (8), 103–106.
- (10) Yuan, Y.; Yuan, J.; Du, H.; Li, L. Pareto Ant Colony Algorithm for Building Life Cycle Energy Consumption Optimization. *Life System Modeling and Intelligent Computing, Pt II* **2010**, *98*, 59–65.
- (11) Zhou, Q.; Zhou, H.; Zhu, Y.; Li, T. Data-driven Solutions for Building Environmental Impact Assessment. In *IEEE 9th International Conference on Semantic Computing*, Feb 07–09, 2015; IEEE Computer Society: Washington, DC, 2015; pp 316–319.
- (12) Song, R. S.; Keller, A. A.; Suh, S. Rapid Life-Cycle Impact Screening Using Artificial Neural Networks. *Environ. Sci. Technol.* **2017**, *51* (18), 10777–10785.
- (13) Suh, S.; Huppes, G. Missing Inventory Estimation Tool using extended Input-Output Analysis. *Int. J. Life Cycle Assess.* **2002**, *7* (3), 134–140.
- (14) Liben-Nowell, D.; Kleinberg, J. The link-prediction problem for social networks. *J. Am. Soc. Inf. Sci. Technol.* **2007**, *58* (7), 1019–1031.
- (15) Zhou, T.; Ren, J.; Medo, M.; Zhang, Y.-C. Bipartite network projection and personal recommendation. In *2011 International Conference on Applied Social Science (ICASS 2011)*, Vol III, 2011; p 489.
- (16) Zhou, T.; Kuscsik, Z.; Liu, J. G.; Medo, M.; Wakeling, J. R.; Zhang, Y. C. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107* (10), 4511–4515.
- (17) Zeng, W.; Shang, M.-S.; Zhang, Q.-M.; Lue, L.; Zhou, T. CAN Dissimilar Users Contribute to Accuracy and Diversity of Personalized Recommendation? *International Journal of Modern Physics C* **2010**, *21* (10), 1217–1227.
- (18) Zhang, Q.-M.; Shang, M.-S.; Zeng, W.; Chen, Y.; Lue, L. Empirical comparison of local structural similarity indices for collaborative-filtering-based recommender systems. *Phys. Procedia* **2010**, *3* (5), 1887–1896.
- (19) Schafer, J. B.; Konstan, J. A.; Riedl, J. E-commerce recommendation applications. *Data Mining and Knowledge Discovery* **2001**, *5* (1–2), 115–153.
- (20) Goldberg, D. S.; Roth, F. P. Assessing experimentally derived interactions in a small world. *Proc. Natl. Acad. Sci. U. S. A.* **2003**, *100* (8), 4372–4376.
- (21) Clauset, A.; Moore, C.; Newman, M. E. J. Hierarchical structure and the prediction of missing links in networks. *Nature* **2008**, *453* (7191), 98–101.
- (22) Suh, S.; Huppes, G. Methods for life cycle inventory of a product. *J. Cleaner Prod.* **2005**, *13* (7), 687–697.
- (23) Souma, W.; Fujiwara, Y.; Aoyama, H. Complex networks and economics. *Phys. A* **2003**, *324* (1–2), 396–401.
- (24) Boccaletti, S.; Latora, V.; Moreno, Y.; Chavez, M.; Hwang, D. U. Complex Networks: Structure and Dynamics. *Complex Systems and Complexity Science* **2007**, *4* (1), 49–92.
- (25) Wang, X. F. Complex networks: Topology, dynamics and synchronization. *Int. J. Bifurcation Chaos Appl. Sci. Eng.* **2002**, *12* (5), 885–916.
- (26) Lue, L.; Zhou, T. Link prediction in complex networks: A survey. *Phys. A* **2011**, *390* (6), 1150–1170.
- (27) Barrat, A.; Barthélemy, M.; Pastor-Satorras, R.; Vespignani, A. The architecture of complex weighted networks. *Proc. Natl. Acad. Sci. U. S. A.* **2004**, *101* (11), 3747–3752.
- (28) Newman, M. E. J. Analysis of weighted networks. *Phys. Rev. E* **2004**, *70*, 5.
- (29) Wernet, G.; Bauer, C.; Steubing, B.; Reinhard, J.; Moreno-Ruiz, E.; Weidema, B. The ecoinvent database version 3 (part I): overview and methodology. *Int. J. Life Cycle Assess.* **2016**, *21* (9), 1218–1230.
- (30) Weidema, B. P.; Bauer, C.; Hirschier, R.; Mutel, C.; Nemecek, T.; Reinhard, J.; Vadenbo, C.; Wernet, G. *Overview and Methodology: Data Quality Guideline for the Ecoinvent Database Version 3*; Swiss Centre for Life Cycle Inventories, 2013.
- (31) Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.
- (32) U.S. Life Cycle Inventory Database. 2012; <https://www.lcacommons.gov/nrel/search>.
- (33) Cooper, J. S.; Kahn, E. Commentary on issues in data quality analysis in life cycle assessment. *Int. J. Life Cycle Assess.* **2012**, *17* (4), 499–503.
- (34) McKone, T. E.; Nazaroff, W. W.; Berck, P.; Auffhammer, M.; Lipman, T.; Torn, M. S.; Masanet, E.; Lobscheid, A.; Santero, N.; Mishra, U.; Barrett, A.; Bomberg, M.; Fingerma, K.; Scown, C.; Strogen, B.; Horvath, A. Grand Challenges for Life-Cycle Assessment of Biofuels. *Environ. Sci. Technol.* **2011**, *45* (5), 1751–1756.
- (35) Wang, M. Q. *REET 1.5-Transportation Fuel-Cycle Model-Vol. 1: Methodology, Development, Use, and Results*; Argonne National Laboratory: Argonne, IL, 1999.

■ NOTE ADDED AFTER ASAP PUBLICATION

This paper published ASAP on April 6, 2018. The interpretation of equation 5 has been corrected. The corrected paper reposted to the Web on April 10, 2018.