2013 IEEE 13th International Conference on Data Mining

A Masking Index for Quantifying Hidden Glitches

Laure Berti-Équille[1,2]
[1]IRD ESPACE DEV, Montpellier, France
laure.berti@ird.fr
[2]Qatar Computing Research Institute, Qatar
lberti@qf.org.qa

Ji Meng Loh
New Jersey Institute of Technology
Dept of Mathematical Sciences
Newark, NJ, USA
loh@njit.edu

Tamraparni Dasu
AT&T Labs - Research
Bedminster, NJ, USA
tamr@research.att.com

***Abstract*—Data glitches are errors in a data set; they are complex entities that often span multiple attributes and records. When they co-occur in data, the presence of one type of glitch can hinder the detection of another type of glitch. This phenomenon is called *masking*. In this paper, we define two important types of masking, and we propose a novel, statistically rigorous indicator called *masking index* for quantifying the hidden glitches in four cases of masking: outliers masked by missing values, outliers masked by duplicates, duplicates masked by missing values, and duplicates masked by outliers.**

The masking index is critical for data quality profiling and data exploration; it enables a user to measure the extent of masking and hence the confidence in the data. In this sense, it is a valuable data quality index for measuring the *true* cleanliness of the data. It is also an objective and quantitative basis for choosing an anomaly detection method that is best suited for the glitches that are present in any given data set. We demonstrate the utility and effectiveness of the masking index by intensive experiments on synthetic and real-world datasets.

Keywords*—*Anomaly detection, data cleaning, duplicate record identification, masking, missing values, outlier detection

I. Introduction

Data glitches are errors in the data that can significantly impact the analysis and conclusions drawn from the data. They occur in a wide variety of ways, ranging from human error (e.g., typos, duplicate entries), to software and hardware problems (e.g., missing values due to transmission failure). As data become more structurally complex and heterogeneous, the gathering, storing and monitoring of data become dependent on intricate systems of interconnected hardware and software. There are numerous opportunities for data to go bad at each of these stages, introducing a daunting variety and quantity of glitches into the data in complex and interrelated patterns.

Financial data streams, communication network data, social data and scientific data, almost all real-world data suffer from missing values, incomplete and distorted values, inconsistent values, duplicate records, and outliers, to mention just a few. These glitches do not occur randomly or in small proportions. They often touch very specific sections of the data, introducing biases into the analysis of the remaining data. The glitches also occur in patterns, as in overloaded network devices with extremely high (outlying) loads that result in outages (missing values). Previous work [1] has addressed and formalized the definition of complex glitches and glitch patterns. Sometimes, one type of glitch makes the other undetectable. For instance, when there are missing values, we might not be able to detect duplicates or true outliers.

When the presence of one type of glitch *masks* another type of glitch and impedes its detection, it can have far-reaching consequences. Masking could result in underestimating the number of glitches and consequently, the cost of cleaning the data. It could also give a false confidence in the results of data analysis. As mentioned earlier, if the masked glitches occur non-randomly in a systematic pattern, they could seriously bias the conclusions drawn from the analysis.

In the past, masking was discussed specifically in the context of outlier detection. Masking, along with the related notion of *swamping* where outliers are duplicated to such an extent that they dominate the distribution and "normal" values become outliers, have been proposed in [2]. In [3], the authors give an intuitive understanding of these effects. Additional references include [4]–[7].

In our work, we focus on masking, but generalize the definition to apply widely to any type of glitch, considerably expanding the scope of previous work beyond outlier detection. We provide a mathematical definition of masking and propose a statistically rigorous method for quantifying masking through a *masking index*.

The masking index is a critical tool for data quality profiling, data exploration, preparation, and mining. It serves two important purposes.

- It enables us to quantify the "hidden" glitches in data and estimate the confidence we have in the results derived from the data.
- It enables us to empirically choose a best glitch detection method when there are multiple glitches in a data set, which is frequently the case in real-world data.

An interesting consequence of masking is that cleaning one set of glitches can reveal (or "unmask") other glitches. For instance, imputing missing values can create "new" duplicate records and outlying values. Or, removing outliers can introduce new duplicates. In such a setting, where there is a need for an iterative approach to data cleaning, the masking index plays an important role in determining the cleanliness of the data and determining the best strategies for cleaning and preparing the data.

Our original contributions can be summarized as follows:

- We propose a general definition of masking that applies widely to any glitch type, including multivariate

1550-4786/13 $31.00 © 2013 IEEE
DOI 10.1109/ICDM.2013.16

glitches, such as outliers and duplicate records and univariate glitches, such as missing values;

- We define two distinct notions of masking, *inner* and *outer* masking, with their respective scope;
- We define a novel, statistically rigorous indicator, called *masking index* for quantifying the extent of masking in the data set;
- We provide a framework for empirically evaluating the masking index in four cases of masking, and
- We propose a method for estimating the masking index in two cases: (1) the case where the ground truth is known, based on simulated data, and (2) the case of real-world data where the ground truth is not available.

Our framework for computing the masking index scales effectively to big data sets. We extract smaller data sets from the entire data set and compute the masking index. By replicating this on several subsets to capture sampling error, we can compute a reliable masking index and thereby choose an appropriate detection method. The conclusions drawn from these smaller data sets are generally applicable to the entire data set.

The rest of the paper is organized as follows. In Section II, we introduce an illustrative example to explain masking. In Section III, we define the problem of masking, introduce the notation, and describe its main features. In addition, we present two different types of masking. In Section IV, we formally define the masking index. We also introduce a theoretical and conceptual formulation of the index that helps us compute it in real world settings where the ground truth is not available.

In Section V, we demonstrate the validity of our approach on synthetic data sets where we control the occurrence of glitches. This allows us to empirically estimate the masking index and study its canonical behavior.

In Section VI, we discuss computing the masking index in real-life scenarios where the ground truth is not known. The theoretical formulations of Section IV combined with resampling from clean parts of the real-world data allow us to compute the masking index. We demonstrate the utility of our contributions on publicly available datasets.

Finally, in Section VII, we discuss existing literature that is relevant to this paper. In Section VIII, we summarize the salient points of the paper and outline future work.

II. An Illustrative Example

To illustrate the complexity of data glitches and the problem of *masking*, consider the data set available at OpenData by Socrata[1] which contains information about Canadian unclaimed bank accounts at branches in the Edmonton area or registered to addresses in the Edmonton area. The web site claims that the total of such abandoned accounts amounts to more than 7 million Canadian dollars.

Here, we consider a subset of 20 records (see Table I). Each record contains the following information about the banking accounts: business or last name (B/LN), first name (FN), balance (BL), address (AD), city (CY), last transaction date (LT), and bank name (BN). A straightforward sum of the unclaimed money in these 20 accounts is $CAN 15,542.10.

However, notice that there are missing values (NULL, _, UNKNOWN, ?? and blanks among others), duplicate records, and outlying or suspicious values in these 20 records. For example, records $x_{18.}$ and $x_{20.}$ appear to be duplicates, with the same transaction dates and very similar locations. The sum of money involved in these two records are identical except for a missing decimal point in $x_{18.}$. This in turn results in a large value of $10,712$ for the balance in cell $x_{18,3}$. Hence what may be outliers in $x_{4,3}$ and $x_{14,3}$ of $CAN 1675.07 and $CAN 1627.5 respectively can be masked, while the possibly legitimate values of $CAN 0.01 in $x_{10,3}, x_{11,3}, x_{12,3}$, and $x_{13,3}$ become swamped. Other glitches may be more subtle. For example, St Albert Trail and McKenney are actually the same location and may mask the fact that $x_{3.}$ and $x_{4.}$ are duplicate accounts.

The process of cleaning the data and computing a total balance is complex. Depending on how duplicates and outliers are treated, e.g., taking the mean account balance or taking the most recent transaction of duplicate accounts, different total balances may be obtained. Figuring out the exact strategy to clean this data set and obtain a realistic total sum of money is beyond the scope of this paper and is part of our future work on iterative cleaning, but this example clearly motivates the need for estimating the number of hidden glitches and understanding how glitch detection can be affected by the masking phenomenon.

III. The Masking Problem

Suppose the data set X is an N by V matrix, with N records and V variables, and that there are K different types of glitches of interest (e.g., missing, outlying values, duplicate records). For each x_{ij} in X, we define a glitch vector $\mathbf{g}_{ij}$ with K elements, where each element $g_{ijk}(k = 1, \ldots, K)$ is 1 if x_{ij} is a glitch of type k, and 0 otherwise. Hence $\mathcal{G}$ is a $N \times V \times K$ array. The array $\mathcal{G}$ represents the true occurrence of glitches in the data X and we refer it as the *ground truth*.

Further, we define $\mathcal{G}'$ to be the array that results from applying glitch detection methods to X. In a world with perfect detection methods, $\mathcal{G} = \mathcal{G}'$, but in reality, the matrix of comparisons between real and detected glitches, $\mathcal{G}' - \mathcal{G}$, contains elements of 0, 1 and -1. An element of 0 means a correct detection (true positive) or a correct non-detection (true negative). An element of 1 means there is a false detection (false positive). An element of -1 means that there is a false non-detection (false negative).

Masking arises when we are not able to detect a glitch due to the presence of another. Non-detection can happen in three ways. First, non-detection may be due to a lack of statistical power of a detection method. Second, the power of the detection method of glitch type k may be reduced by the presence of glitch type k'. The size of both these effects depend on the specific detection method used. The third possible cause

[1] Data set from OpenData by Socrata retrieved on March 26, 2013: https://opendata.socrata.com/Government/Unclaimed-bank-accounts/

TABLE I. A SUBSET OF 20 RECORDS TAKEN FROM THE "UNCLAIMED BANK ACCOUNTS" DATA SET FROM OPENDATA BY SOCRATA[1] WITH EXAMPLES OF MISSING VALUES, DUPLICATES AND OUTLYING VALUES

	$x_{.1}$	$x_{.2}$	$x_{.3}$	$x_{.4}$	$x_{.5}$	$x_{.6}$	$x_{.7}$
x_{ij}	**B/LN**	**FN**	**BL** ($CAN)	**AD**	**CY**	**LT**	**BN**
$x_{1.}$	CANADIAN	BERND/CANADIAN JANE	30	BOX 36 SITE 6 RR 2	THORSBY AB	10/22/1994	BANK OF MONTREAL
$x_{2.}$	CANADIAN	BERND/CANADIAN JANE	30	BOX 36 SITE 6 RR 2	THORSBY AB	10/19/1994	BANK OF MONTREAL
$x_{3.}$	TRUST AC		278	MCKENNEY	St. Albert	11/30/1993	
$x_{4.}$	TRUST AM		1675.07	ST ALBERT TRAIL	St. Albert	11/30/1993	TORONTO-DOMINION BANK
$x_{5.}$	BRUNO	DAKOTA H	5.02	_	_	_	BANK OF NOVA SCOTIA
$x_{6.}$	BRUNO	DANIEL S	5.02		M1M 1M1	10/30/1992	BANK OF NOVA SCOTIA
$x_{7.}$	BRUNO	GRANT C	5.02	UNKNOWN	UNKNOWN	10/30/1992	BANK OF NOVA SCOTIA
$x_{8.}$	BRODERICK	MARGARET	122.91	20 OAK ST	Sherwood Park	12/21/1995	CAN IMPERIAL BANK OF COM
$x_{9.}$	BRODERICK	MARGARET	107.88	20 OAK ST	Sherwood Park	12/22/1995	CAN IMPERIAL BANK OF COM
$x_{10.}$	MURPHY	DOYLE	0.01	34 WOODVALE	AB	10/07/1992	BANK OF NOVA SCOTIA
$x_{11.}$	MURPHY	MEGAN	0.01	34 WOODVALE	AB	10/07/1992	BANK OF NOVA SCOTIA
$x_{12.}$	QUINTAL	DANI	0.01	RR 1	CALAHOO AB	05/09/1991	BANK OF NOVA SCOTIA
$x_{13.}$	QUINTAL	MEGAN	0.01	165 Woodbuffalo way	Ft McMurray AB	05/09/1991	BANK OF NOVA SCOTIA
$x_{14.}$	YOUNG	MUSICIANS ACADAMY	1627.5	??	??	01/03/1975	ROYAL BANK OF CANADA
$x_{15.}$	YOUNG	MUSICIANS ACADAMY	76.06	NULL	NULL	01/04/1975	ROYAL BANK OF CANADA
$x_{16.}$	ZITTLAW	EDWARD	410.27			07/02/1988	ROYAL BANK OF CANADA
$x_{17.}$	ZITTLAW	EDWARD	341.53			02/07/1988	ROYAL BANK OF CANADA
$x_{18.}$		BUSCH*WILLIAM*JACKSON	10712	JASPER AVE NW	EDMONTON	04/18/1986	TORONTO-DOMINION BANK
$x_{19.}$	BUSH	WILLIAM	8.66			08/16/1986	MONTREAL TRUST COMP OF CA
$x_{20.}$	WILLIAM	*BUSCH*J	107.12	10230 JASPER AVE	EDMONTON	04/18/1986	TORONTO-DOMINION BANK
		SUM	15,542.10				

is a direct effect of glitch type k' on glitch type k, independent of the detection method used. We describe these three effects in the following subsections.

A. Power of Detection

Glitch detection methods vary in their ability to detect glitches. One measure of performance of a detection method m is the *statistical power*, $\pi_m \in [0, 1]$. Traditionally, power is defined in the context of clean data. Suppose the data is completely clean except for the single glitch of type k in x_{ij} that the method has to detect.

Definition 1. *The statistical power of method m for detecting glitch type k is the probability that the glitch is detected. It is given by:*

$$\pi_{m,k} = P(\mathcal{G}'_{ijk} = 1 | \mathcal{G}_{ijk} = 1).$$

Except for cases where the detection is absolute (e.g., the detection of missing values), typically, $\pi_{m,k}$ will be close to but less than 1, indicating a good but imperfect method.

A detection method generally has power $0 < \pi_{m,k} < 1$ even when applied to clean data due to random error. However, the presence of other types of glitches may interact with the detection method, resulting in a change in power. Let $\epsilon_{m,k,k'}$ be the change in the power of method m under the influence of other glitches of type k'. We will assume that $0 \leq \epsilon_{m,k,k'} \leq \pi_{m,k}$ though it might be possible for one type of glitch to improve power of detection of another type of glitch. Intuitively, a detection method that is not affected or only slightly affected by the presence of other data glitches, ***i.e.***, $\epsilon_{m,k,k'} \approx 0$, is said to be robust to data glitches of type k'. The altered power is given by:

$$\begin{aligned} \pi_{m,k,k'} &= P(\mathcal{G}'_{ijk} = 1 | \mathcal{G}_{i'lk'} = 1 \wedge \mathcal{G}_{ijk} = 1, i \neq i', l = ., j') \\ &= \pi_{m,k} - \epsilon_{m,k,k'} \end{aligned}$$

Therefore,

Definition 2. *The robustness of method m in detecting glitches of type k to the presence of glitches of type k' is the ratio of the altered power $\pi_{m,k,k'}$ in the presence of glitches of type k', to $\pi_{m,k}$, the power of detection without glitches of type k'. That is,*

$$\rho_{m,k,k'} = \frac{\pi_{m,k,k'}}{\pi_{m,k}}.$$

To simplify notation, we will henceforth drop the subscript m in all expressions without losing clarity or generality, and refer to these quantities as $\pi_k, \epsilon_{k,k'}$ and $\rho_{k,k'}$ instead, for a given method.

B. Types of Masking

There is a fundamental connection between masking and power. Power is a measure of the ability to detect, while masking is exactly the opposite. The connection between the two can be expressed as:

$$\Pi = 1 - \mathcal{M} \quad (1)$$

where Π is power of detection and $\mathcal{M}$, the probability of masking. We explain further in the following sections.

First, it is useful to make a distinction between two basic types of masking. A glitch of type k' at the data cell x_{ij} could mask a glitch of type k in the same record i, or in a different record i'. This leads to the notion of two types of masking, *inner* and *outer masking*.

Before defining them, we illustrate them by considering the schematic depiction in Figure 1. First consider the case where only glitches of type k (blue dots) are present. Most of them are detected in the absence of glitches of type k' (pink diamonds) as indicated by the large brace on the left. The bottom record is a false positive. Furthermore, there are some glitches of type k that are not detected if the detection method's power is less than one. These false negatives are highlighted by an ellipse around them in the top left side of the figure.

Now suppose that we introduce glitches of type k'. We fail to detect some of the glitches of type k that we could detect earlier, as shown by dashed lines. The braces on the left highlight the two possibilities. When glitches of type k'

are present in the same record i as glitch k, and we are unable to detect a glitch of type k that we could detect earlier, we call this phenomenon *inner masking*. When the glitch k' is in another record, and we still fail to detect glitch k, it is *outer masking*. We now define these two notions of masking below.

Definition 3. *Inner masking is the phenomenon where the detection of a glitch of type k' prevents the detection of a glitch of type k in the same cell or record. That is, $\mathcal{G}'_{ijk'} = 1 \implies \mathcal{G}'_{ijk} = 0$ even though $\mathcal{G}_{ijk} = 1$. The glitches could be of the same type or of different types. The extent of inner masking is given by:*

$$\mathcal{M}(k/k')|_{inner} = P(\mathcal{G}'_{ijk} = 0 | \mathcal{G}'_{ilk'} = 1 \wedge \mathcal{G}_{ijk} = 1) \quad (2)$$

with $l = ., j$.

In the probability statement above, we made a simplifying assumption:

Assumption 1: If glitches of type k and k' are present on the same record i, then k is always masked.

For example, outliers will always be masked if there is a missing value present in the cell. Default values may systematically cover up the presence of missing values. Under this assumption, we have a phenomenon of *glitch dominance*. More specifically, we say that a glitch type k' always dominates glitch type k if

$$\mathcal{G}'_{ijk'} = 1 \implies \mathcal{G}'_{ij'k} = 0, \quad \forall i, j, j',$$

so that

$$\mathrm{P}(\mathcal{G}'_{ijk} = 1 | \mathcal{G}'_{ij'k'} = 1 \wedge \mathcal{G}_{ijk} = 1) = 0, \quad \forall i, j, j'.$$

Relaxing Assumption 1 requires further splitting the expression in Equation 2 by the conditional probability of detecting k in the presence of k' in the same record.

In addition to affecting glitch detection in the same record, glitches could interfere with the detection of glitches in other records. We define this notion of *outer masking* below.

Definition 4. *Outer masking is the phenomenon where the occurrence of a glitch of type k' in one record, hinders the detection of glitches of type k in other records. The glitches could be of the same type or of different types.*

$$\mathcal{M}(k/k')|_{outer} = P(\mathcal{G}'_{ijk} = 0 | \mathcal{G}_{i'lk'} = 1 \wedge \mathcal{G}_{ijk} = 1) \quad (3)$$

with $i \neq i'$ and $l = ., j'$.

Again, the probability statement involves a simplifying assumption:

Assumption 2: If a glitch of type k' is present, we will always detect it. That is, there exists a method m such that $\pi_{m,k'} = 1$ and hence we can assume that $\mathcal{G}'_{ijk'} = \mathcal{G}_{ijk'}, \forall i, j$ for method m.

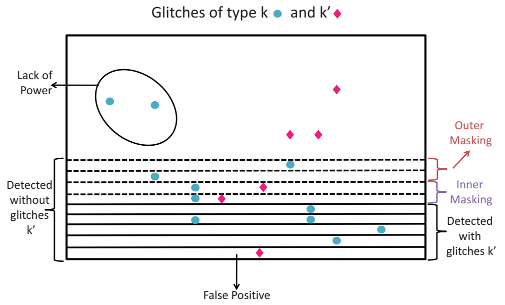

Glitches of type k ● and k' ♦
Lack of Power
Outer Masking
Inner Masking
Detected without glitches k'
Detected with glitches k'
False Positive

Fig. 1. Inner and outer masking: Glitches of type k (blue dots) that are detectable become undetectable (dashed lines) with the introduction of glitches of type k' (pink diamonds). If k' and k are in the same record i, it is *inner masking*, and if they are in different records it is *outer masking*. In addition, a method of detection would also result in false positives (bottom record) and false negatives (shown inside ellipse on top left of the figure).

Missing values are an excellent example of glitches that could be detected with certainty. Relaxing Assumption 2 requires an additional step in Equation 3 of conditioning on the power of detection k'. Note that we do not require these assumptions to do the empirical studies. Outer masking is related to the change in power of detecting glitches of type k due to the presence of glitches of type k' in other records, and can be motivated using the concept of *robustness* of Definition 2 as follows. Suppose that we use a method with power π_k of detecting glitches of type k in the absence of glitches of type k'. Suppose further that when glitches of type k' are introduced, the power changes. Some glitches of type k that occur on the same record as glitches of type k' disappear due to inner masking. Others of type k disappear due to outer masking by glitches of type k'. From Definition 2 the changed power is:

$$\pi_{k,k'} = (\pi_k - \epsilon_{k,k'}) = \rho_{k,k'} \pi_k.$$

Therefore the probability that a glitch of type k will not be detected (outer masked), using the fundamental relationship in Equation 1 is given by:

$$\begin{aligned} \mathcal{M}(k/k')|_{outer} &= 1 - \pi_{k,k'} \\ &= 1 - \rho_{k,k'} \pi_k. \end{aligned} \quad (4)$$

IV. The Masking Index

In this section, we construct an index to quantify the masking effect of glitches based on the probability of glitch detection in the presence of other glitches.

Definition 5. *The masking index of glitch type k with respect to glitch type k' is defined as the probability that the presence of a glitch of type k is masked by the presence of glitches of type k':*

$$\mathcal{M}_{k/k'} = P(\mathcal{G}'_{ijk} = 0 | \mathcal{G}_{ijk} = 1 \wedge \mathcal{G}_{i'l'k'} \neq 0)$$

with $l' = ., j'$.

We can formulate the masking index in an alternate way to aid computation. A glitch is detected (not masked) if it is not inner masked and not outer masked, *i.e.*,

$$1 - \mathcal{M}_{k/k'} = A \times B,$$

where from Equations 2 and 3,

$$A = 1 - \mathcal{M}(k/k')|_{inner}$$

and

$$B = 1 - \mathcal{M}(k/k')|_{outer}$$

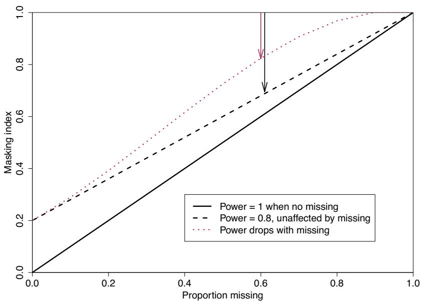


Fig. 2. Masking Index: The X-axis depicts the proportion of missing values. The Y-axis represents the masking index or the probability of non-detection of outliers due to the presence of missing values. Each plot line corresponds to the masking index of a given method. The solid line represents a detection method with power 1, and the dashed line represents a detection method with power less than 1 but robustness 1 (not outer masked by missing). The difference between the dotted and dashed lines represents the additional effect due to outer masking.

and therefore the masking index is

$$\mathcal{M}_{k/k'} = 1 - A \times B.$$

The above discussion can be represented by means of Figure 2. The X-axis depicts the proportion of glitch type k', in this case the varying proportion of missing values. The Y-axis represents the masking index or the probability of non-detection of glitch of type k, (e.g., outliers), due to the presence of glitches of type k' (e.g., missing values). Each plotted line corresponds to the masking index of a given method. Therefore, according to Equation 1, if we drop a perpendicular line from the top of the plot to the any of the curves (depicted by arrows in the figure), the length of that line would represent the power of the method.

In this figure, the solid line represents a detection method of glitch type k with power 1, so the masking index is solely determined by the prevalence of the masking effect of glitch type k'. The dashed line represents the case where the detection method has power less than 1 (we used 0.8 in the figure), but whose performance is unaffected by the presence of type k' glitches, *i.e.*, its robustness is $\rho_{k,k'} = 1$. A method whose performance in detecting type k glitches is affected (impaired) by the presence of type k' glitches, with $\rho_{k,k'} < 1$, is represented by the dotted line in Figure 2.

The difference between the dotted and dashed lines shows the size of the effect on the performance of the detection method due to the prevalence of type k' glitches. Note that the non-robust method loses more power as more glitches get masked compared to the other two robust methods. We study specific cases of masking in the next section instantiating A and B from the previous definition.

A. *Specific Cases of Masking*

We now discuss the conceptual formulations of the terms A and B for different instances of masking. The formulation depends on the nature of interaction between the two types of glitches considered. The conceptual formulations are important for computing the masking index in real-world data sets where the ground truth is not known.

When the ground truth is known, the quantities A (inner masking) and B (outer masking) can be estimated empirically. Glitches of type k are flagged, with and without the presence of glitches of type k', and then compared with the ground truth. By controlling the proportion of glitches of type k and k', we can understand the behavior of the masking index $\mathcal{M}_{k/k'}$.

We defer a detailed discussion of the experiments and data simulation to Section V and Section VI.

Masking of numeric outliers by missing values

Let the outlier detection method O have power π_O and robustness $\rho_{O/M}$ to missing values. Suppose that the proportion p_M of missing values is scattered randomly throughout the data set, so that the probability that an outlier and missing value occur together leading to masking of the outlier is given by p_M. Then, the masking index is given by

$$\mathcal{M}_{O/M} = 1 - (1 - p_M)\pi_O \rho_{O/M}. \tag{5}$$

Here, $A = 1 - p_M$ represents the probability that an outlier is not inner masked by a missing value. This probability is the same as that for any specific cell x_{ij} in the data set. The term $B = \pi_O \rho_{O/M}$ represents the newly changed power of the outlier detection method in the presence of missing values.

Masking of numeric outliers by duplicates

An outlier can be masked by duplicates if it is duplicated so many times that the outlier value becomes part of the normal portion of the data distribution as determined by the detection method.

Suppose records are randomly duplicated (either exactly or with slight errors so they become approximate duplicates). Suppose further that there is a process that generates these duplicate records, such that there is a probability p_d that the record is duplicated d times ($d = 1, 2, \dots$). Thus $\sum_d p_d = 1$.

Clearly, for a moderately sized data set, if an outlying value is duplicated two times, say, it is unlikely to affect the distribution of values much. Conceptually, there is some threshold K: if the outlying value is duplicated at least K times, these K values overwhelm the rest of the distribution. The probability of this NOT happening (and hence not inner masked) is $\sum_{d<K} p_d$. Even if $d < K$, the power of detection method may still be affected, with the new power given by $\pi_O \rho_{O/D}$, where $\rho_{O/D}$ is the robustness. This is related to outer masking. This suggests that a masking index for the masking of outliers by duplicates is of the form

$$\mathcal{M}_{O/D} = 1 - \left(\sum_{d<K} p_d \right) \pi_O \rho_{O/D}. \tag{6}$$

The difficulty in applying this expression, however, is that K and the p_d's are unknown and not easily inferred.

Masking of duplicates by missing values

Suppose that a record $x_{i,.}$ is duplicated d times, with duplicate records denoted by $x^1_{i,.}, x^2_{i,.}, \ldots x^d_{i,.}$. The record $x_{i,.}$ is not identified as a duplicate if $x_{i,.}$ contains missing values, or if each of $x^1_{i,.}, x^2_{i,.}, \ldots x^d_{i,.}$ contains missing values. Hence the probability that x_i is not inner masked is $(1-p_M)(1-p_M^d)$. With the power and robustness given by π_{Dd} and $\rho_{Dd/M}$ respectively (the subscript Dd represents duplication with d additional records), the masking index conditioned on a record with d duplicates is

$$\mathcal{M}_{Dd/M} = 1 - (1-p_M)(1-p_M^d)\pi_{Dd}\rho_{Dd/M}. \qquad (7)$$

Note that both π_{Dd} and $\rho_{Dd/M}$ may depend on the actual number of duplicate records d, since duplicate detection might be easier with more duplicate records, for example.

The masking index for any duplicate record (duplicated any number of times) is then

$$\mathcal{M}_{D/M} = 1 - \sum_d p_d(1-p_M)(1-p_M^d)\pi_{Dd}\rho_{Dd/M} \qquad (8)$$

where the second term is a weighted sum with weights given by p_d, the probability that a record is duplicated d times.

Masking of duplicates by outliers

Similarly, outliers can mask duplicates in the same way as missing values, making the records different enough to be undetected as a duplicate. The masking index is given by

$$\mathcal{M}_{D/O} = 1 - \sum_d p_d(1-p_O)(1-p_O^d)\pi_{Dd}\rho_{Dd/O} \qquad (9)$$

where p_O is the probability that any value x_{ij} is an outlier, and π_{Dd} is the power of the duplicate detection method for detecting d duplicates and $\rho_{Dd/O}$ the robustness to outliers.

In the rest of this paper, we study the masking index empirically, first using synthetically generated data (Section V) and second with real-world data sets (Section VI). With synthetically generated data sets, the occurrence and amount of glitches can be controlled. Knowing the ground truth allows us to unambiguously quantify the amount of masking as well as the different contributions of inner and outer masking under various scenarios. With the real-world data sets, we quantify the degree of masking that exists by estimating the masking index with the ground truth unknown. Since the masking index is computed with respect to a specific method, we also illustrate how we might use the masking index to choose a detection method that is least affected by masking, out of several choices of detection methods.

V. Masking Index through Simulations

As seen from Definitions 3 and 4, the masking index is defined with respect to a pair of glitches of types k and k'. We ran experiments to study two cases of masking effect: (1) masking of missing values (type k') on outliers (type k), and (2) masking of duplicates (type k') on outlier (type k) detection. Simulations provide a controlled environment to study the canonical behavior of the masking index. All the experiments were conducted using the R statistical package for data generation, outlier and duplicate detection, and for computing the masking index.

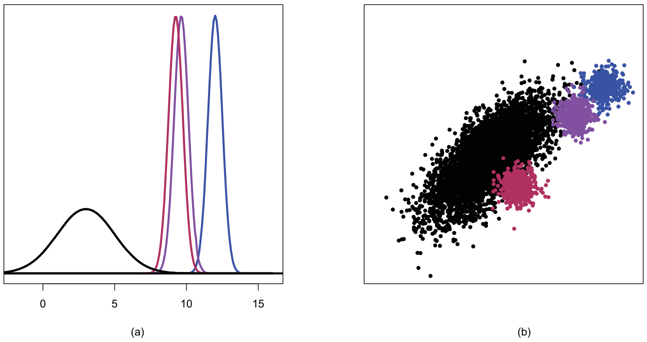


Fig. 3. Baseline distributions (black) and corrupting outlying distributions: (a) Univariate N$(3, 4)$ baseline, with outlier distributions of means of 12, 9.64 and 9.26, and variance 0.25; (b) Bivariate normal (baseline, as described in text) with corrupting distributions of independent bivariate normals with variance 0.1 and means $(4, 2.5)$, $(3, 1.5)$ and $(1, -1)$.

We started by creating a *baseline data set* of 5000 records, each with 4 values – a random string, a univariate normal variable and the two components of a bivariate normal variable. The univariate normal was N$(3, 4)$. The bivariate normal was specified by:

$$\mathrm{N}\left(\left(\begin{array}{c}0\\0\end{array}\right), \left(\begin{array}{cc}1 & 0.7\\0.7 & 1\end{array}\right)\right).$$

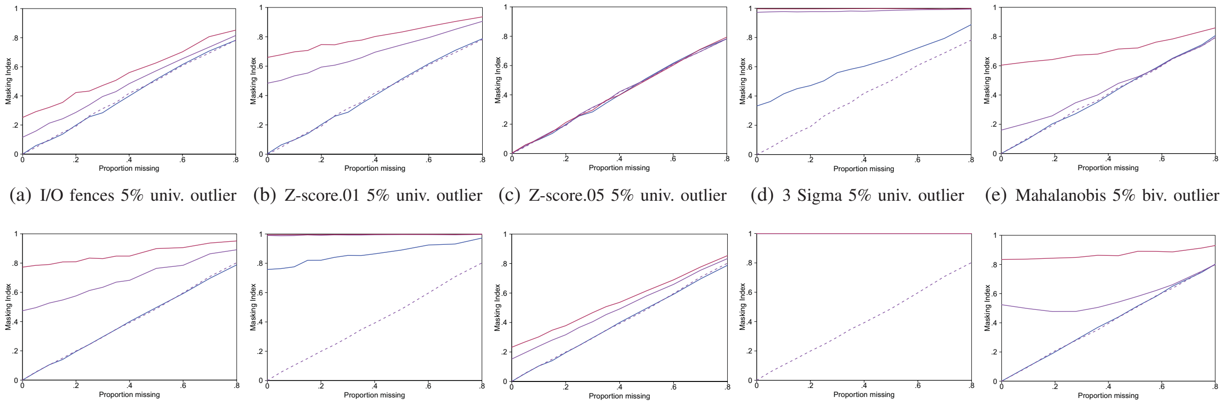
A. Missing Values and Outliers

To study the masking of outliers by missing values, we generated *corrupt data sets* by injecting outliers and missing values in varying proportions.

Generating Outliers: The outlier values for the univariate normal variable in the baseline data set were generated by drawing from a different univariate distribution, and similarly, from a separate bivariate normal distribution for corrupting the bivariate normal variable.

- For the univariate normal, the outlier distributions used were normal with variance 0.25, and means of 12, 9.64 and 9.26. The outlier distributions are chosen to be increasingly difficult to detect, resulting in a decrease in the power of the detection methods.
- For the bivariate normal variables, the outlier distributions were independent bivariate normals with variance 0.1 and means $(4, 2.5), (3, 1.5)$ and $(1, -1)$. These outliers distributions are detected with decreasing power by any specific detection method as they become increasingly similar to the baseline distribution.
- We injected 5% and 10% of outlier values, from each outlier distribution in turn.

Figure 3 shows a picture of the univariate population and outlier densities and sample realizations from the bivariate distributions we used. The baseline distribution is shown in black.



(a) I/O fences 5% univ. outlier (b) Z-score.01 5% univ. outlier (c) Z-score.05 5% univ. outlier (d) 3 Sigma 5% univ. outlier (e) Mahalanobis 5% biv. outlier

(f) I/O fences 10% univ. outlier (g) Z-score.01 10% univ. outlier (h) Z-score.05 10% univ. outlier (i) 3 Sigma 10% univ. outlier (j) Mahalanobis 10% biv. outlier

Fig. 4. Masking index for outlier detection in the presence of missing values $\mathcal{M}_{O/M}$. The X-axis denotes the proportion of missing values. The Y-axis is the masking index of Equation 5, estimated empirically from the simulated data sets where the ground truth is known. The linear trend in the masking index is due to inner masking (dashed line).

Generating Missing Values: To create missing values, we removed data at random, with the proportion of missing values ranging from 0.1 to 0.8. The original baseline data set corresponds to missing value proportion = 0, and at the other extreme, all the data are missing when missing value proportion = 1.
Outlier Detection: For outlier detection, we used the following methods defined in [7]:

- Four well-known univariate methods: (1) z-score with p=.01 and (2) z-score with p=.05, (3) inner ($Q1 \pm 1.5 * IQR$) and outer fences ($Q3 \pm 1.5 * IQR$) noted I/O fences method, and (4) 3σ on the continuous variables;
- Two multivariate methods based on: (1) Mahalanobis distance (see [8] for details) and (2) Jackknife distance (see [9] for details).

Computing the Masking Index $\mathcal{M}_{O/M}$: We computed the masking index very simply: by counting how many outliers were detected before and after the injection of missing values. Note that while it is obvious that some outliers are explicitly knocked off by the missing values (inner masking), other values that were outliers in the baseline data were not flagged as outliers in the corrupted data set even though there was no missing value present at the record containing the former outlier. This is an example of outer masking.

Figure 4 shows the masking index for outlier detection in the presence of missing values, $\mathcal{M}_{O/M}$. The first four panels from (a) to (d) correspond to the univariate outlier detection methods using the inner and outer fences, z-score with p=.01, z-score with p=.05 and 3 sigma methods and the fifth panel (e) corresponds to bivariate outlier detection with the Mahalanobis method. The X-axis denotes proportion of missing values. The Y-axis is the masking index of Equation 5, estimated empirically from the simulated data sets where the ground truth is known. Since Jackknife bivariate outlier detection method behaves very similarly to Mahalanobis method in both cases of outlier injection, we did not report the figures. In the four univariate cases, the baseline data set was contaminated with 5% data from an outlying distribution shown in Figure 3(a). In panel (e) 5% of the data are injected with values from the outlying bivariate distributions represented in Figure 3(b). In the second line of panels from (f) to (j), 10% of the data are outliers.

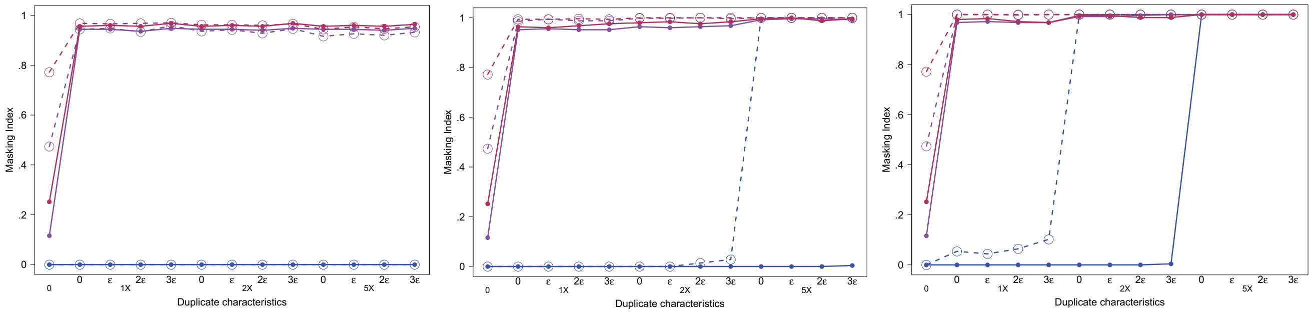
The linear trend in the masking index is due to inner masking (dashed line). Each solid curve shows the masking index corresponding to a different outlying distribution as mentioned in Section V-A with variance 0.25, and means of 12 (blue line), 9.64 (purple line), and 9.26 (red line). According to the fundamental relationship between the masking index and power defined in Equation 1, the power can be read from the top of the plot down to the curves. Although the same detection method is used within each panel in Figure 4, the power is different due to the different alternatives (different means of the outlying distributions). As expected, the masking effect on outlier detection increases as the proportion of missing values increases. This is a consequence of inner masking. The amount of inner masking is almost the same for all the outlying distributions and in each panel; we show it for one distribution using a dashed line.

The result confirms our intuition that inner masking is linear in the proportion of missing values. Note also the difference in the masking index (and hence the power) between Figures 4(a)-(d) and 4(f)-(i) caused by the different amount of outlying values (from 5 to 10%). The roughly straight lines in Figure 4 suggest that the methods are not unduly affected by the amount of missing values. As the amount of inner masking increases with the proportion of missing values (the dashed line), the amount of non-detection due to a lack of power (the gap between the solid curve and the dashed line) decreases.

B. Outliers and Duplicates

To study the masking index for outliers (type k) in the presence of duplicates (type k'), we used the same baseline data set as discussed in Section V-A as well as those data sets with 5% and 10% outliers injected, and introduced duplicates to them.
Generating Duplicates: To inject duplicates,

- We first split the baseline data sets into non-outlying records and outlying records, creating D_N and D_O respectively.



(a) 0% duplicates are outliers (I/O fences method) (b) 50% duplicates are outliers (I/O fences method) (c) All duplicates are outliers (I/O fences method)

Fig. 5. Masking index for outlier detection in the presence of duplicates $\mathcal{M}_{O/D}$. The three panels correspond to where the duplicates come from, 0%, 50% and 100% from the outliers, respectively. The amount of outliers, 5% or 10%, are indicated by solid lines/dots and dashed lines/clear dots respectively. Different colors correspond to the outlying univariate distributions shown in Figure 3(a). Within each panel, the leftmost point shows the masking index when there are no duplicates. The next set of 4 points are 1× duplication (one exact, and 3 approximate) followed by 2× and 5× duplication.

- We created 3 sets of building block duplicate records by drawing 100%, 50% or 0% from the outlying records D_O and the rest from the non-outlying records D_N.
- We replicated these building block duplicate sets 1, 2 and 5 times, yielding the final duplicate sets.
- The final duplicate sets were then added to the baseline data sets in turn.
- We also considered exact and approximate duplication. For approximate duplication there is an intermediate step applied to the duplicate sets before they are added to the baseline data sets: random values drawn from a uniform distribution with endpoints $\pm\epsilon$, $\pm 2\epsilon$ and $\pm 3\epsilon$ are added to the variable of interest (v). We chose ϵ to be $0.05\sigma_v$, where σ_v is the standard deviation of the variable v.

By this process, we added exact and approximate duplicate records of size ranging from 250 to 2500 to the original baseline data sets of size 5000 while also varying the proportion of duplicates that themselves contain outliers.

Duplicate detection: The duplicates were detected based on 4 distance-based detection methods for strings (see [10] for details): Jaro, JaroWinkler, Jaccard coefficient, and Cosine similarity measures and based on Euclidean distance for numeric variables.

Computing the Masking Index $\mathcal{M}_{O/D}$: Since each experiment consisted of random error injection or selection (with various seeds) as described, we replicated the process of each corruption and detection method 5 times, to report and average our results. Then, for each run and each corrupted version of each data set, we applied again the detection methods to compute its masking index with respect to the other type of glitch considered.

Figure 5 shows plots from the experiment with panels (a), (b) and (c) corresponding to duplicates being generated only from non-outlying records (0%), 50% from outlying records and 100% from outlying records. The colors of the dots and lines correspond to the respective outlying distributions used (with different means). The solid lines and dots are for 5% of the data injected with outliers while the dashed lines and clear dots are for 10% outliers. We report results only for the inner fences outlier detection method. The X-axis shows the characteristics of the duplication used in the experiment. The leftmost points of the curves show the masking index without any duplicates. This is followed by 3 groups of 4 points each, labelled “1X”, “2X”, “5X”, representing the cases where the records chosen for duplication are replicated 1, 2 or 5 times. Within each group of points, we have either exact duplicates, or duplicates shifted randomly by different amounts, labelled “0” (for exact), “ϵ”, “2ϵ”, “3ϵ”. Note that there is not a strict order for the duplication characteristics. We join the dots with lines so that the reader can more easily make out the differences between the dots.

We find that for outlying distributions that are very different from the population (blue), duplication of the non-outliers has minimal effect on outlier detection (panels (a) and (b) for 0% and 50% non-outliers duplicated). However, with 100% of duplicates generated from outliers (panel (c)), the masking index becomes high when the number of duplicates is large. With the other outlying distributions with means closer to that of the population, masking becomes very high even with moderate amounts of duplication. There is little difference between exact and approximate duplication, although this might be due to our use of a small value for ϵ.

VI. Masking Index for Real-World Data

We use two publicly available real-world data sets, one on Internet advertisements and one much larger data set on mean sea level differences to:

- Demonstrate the estimation of the masking index when the ground truth is not known, and
- Use the masking index to choose a detection method that is least affected by masking.

For these data sets, there are missing values and no duplicates, and we consider the masking of missing values (type k') on the detection of outliers (type k). Data sets are described in Section VI-B and C. The results are synthesized in Table II.

A. Estimating the Masking Index

Since the data already have missing values, outlier detection is already masked, *i.e.*, outliers cannot be detected at the

TABLE II. Masking Index Estimates for the Real-World Data Sets

					Outlier Detection Methods			
Data Set	*Data Set Description*	*Number of Glitches*			3σ	Inner/Outer Fences	z-score.05	z-score.01
Internet	3,279 records	Missing: 23 (0.7%)	Duplicate detected: 0 (0%)	Outliers detected:	25	33	30	50
Ads	3 continuous variables			$\mathcal{M}_{O/M}$	0.2	0.17	0.21	0.22
Mean Sea	826,000 records	Missing: 13,2160 (16%)	Duplicate detected: 0 (0%)	Outliers detected:	14520	54834	32717	21776
Level Data	1 continuous variable			$\mathcal{M}_{O/M}$	0.62	0.69	0.53	0.51

missing values (inner masking) and the power of the detection methods are already affected by the missing values (outer masking). Thus, to estimate the masking index, we use the following approach:

Step (1) We use the theoretical formulation given in Section IV-A for the masking index, which requires estimation of the power π_O and the robustness $\rho_{O/M}$. The proportion of missing values p_M is also required, but this is, of course, easily obtained.

Step (2) Given multiple methods of detection, we select one method to detect outliers, and treat it as the ground truth. With this specification of the ground truth, we estimate the power and robustness of the other methods, using steps (3) and (4) below. We rotate the role of the methods, so that each method gets to be treated as the ground truth.

Step (3) In order to estimate the power, we divide the data (D) into clean data set (D_{clean}) with no missing values, and dirty (D_{dirty}). Due to inner masking, there are no outliers detected in D_{dirty}. The power of each method is obtained by comparing the number of outliers it detects in D_{clean} with the number of outliers of the ground truth.

Step (4) The robustness of a detection method is found by treating the clean D_{clean} as the complete data and injecting a proportion p_M of missing values to it, creating D_M. The choice of value of p_M is the actual amount of missing values in the original data D. By injecting missing values to the clean D_{clean} we are re-creating the scenario of having proportion p_M of missing values in the original data set D. By comparing the detections in D_{clean} with those in D_M, the change in power and hence the robustness at the observed level of missing values can be estimated.

Step (5) The injection of missing values is repeated multiple times to reduce sampling error and obtain more stable estimates of the robustness. An estimate of the masking index is then obtained using Equation 5. Finally, we take the mean of all the masking indices obtained from the rotation of detection methods used for the ground truth, to obtain the masking index estimate of each method.

Identifying the ground truth can also be done using a voting mechanism. To estimate the masking index of one method, we use the other methods to determine the ground truth – if more methods agree, there is greater confidence that a glitch is real. The procedure described above can be considered a special case where a single method does the voting.

B. Internet Advertisements Data

The Internet advertisements data is described in [11] and is available at the UCI[2]. It contains 3,279 instances representing a set of possible advertisements on Internet Web pages. The features encode the geometry of the advertisement image (if available), specifically the *height, width* and *aspect ratio*. Here, we focus on just the aspect ratio. As presented in Table II, the proportion of missing is small, about 0.7%. This data set has no duplicate record. We used four outlier detection methods: 3σ, inner/outer fences, z-score 0.05 and z-score 0.01. For these methods, the average estimated statistical power is 0.84, 0.84, 0.88 and 0.88 respectively. The corresponding estimated mean robustness values are 0.96, 1, 0.9 and 0.89, based on 100 independent replications of data created from injecting missing values to the clean portion of the data. Using Equation 5, the corresponding estimated masking indices are 0.2, 0.17, 0.21, and 0.22, suggesting that for this data set, the inner/outer fences method is “best” in terms of having the least amount of outliers masked by the missing values.

C. Mean Sea Level Data

The mean seal level data set is extracted from the Permanent Service for Mean Sea Level[3] and contains about 826,000 tuples describing the average change in sea level from tide gauges and bottom pressure recorders all around the world. The data goes back to the 1800s. This data set has more missing values than the Internet advertisements data set, with about 16% missing values. It has no duplicates as well. However, probably due to the extremely large data size, we find that the robustness of all the four detection methods are high, at around 0.998 with little difference between the methods. The ranking of the masking indices was thus determined by the relative power of the methods. The estimated power was 0.59, 0.57, 0.67, 0.68, for the 3σ, inner/outer fences, z-score 0.05 and z-score 0.01 methods, yielding masking indices 0.62, 0.69, 0.53, and 0.51, so that the z-score 0.01 method has the lowest masking index for this data set.

D. Discussion

In this paper, we assume that the glitch matrix can be computed off-line. Many other anomaly detection methods can be used to characterize the *dirtiness* of the data set (and consequently, augment the size of the glitch matrix and the overall computing time). In the experiments on the two real-world data sets, we have demonstrated that the masking index is an effective method for determining a glitch detection method that is least affected by masking. Along with other data quality metrics, the masking index plays a critical role in the selection of data cleaning strategies, particularly in the case of iterative cleaning where it can be used to determine a stopping criterion for iteration. Future work will be devoted to the study of the efficiency, complexity, and scalability of our approach with expanding the set of anomaly detection methods.

[2] http://archive.ics.uci.edu/ml/datasets/Internet+Advertisements

[3] Permanent Service for Mean Sea Level – PSMSL: http://www.psmsl.org/

VII. Related Work

It is common to use data for constructing models that represent real-world behavior in compact and aggregated forms (e.g., formulas, charts, statistical classification or regression models, etc.). These summaries allow the decision maker to understand and analyze certain phenomena and behaviors. Model complexity and reliability may be significantly affected by data cleaning and preparation processes and their resulting data quality. The link between data quality and decision model correctness, which has been explored in a variety of studies (e.g., [12]), is still very complex and difficult to assess.

Removing anomalies and noisy data is an important goal of data cleaning because noise and errors hinder most types of data analysis. Most existing data cleaning methods from database research, data mining and statistics literature focus on removing noise as the result of low-level data errors from an imperfect data collection process [13], but the masking effect of a conjunction of anomalies can significantly bias data preparation and hinder analysis.

To the best of our knowledge, there is little work focused on iterative cleaning to efficiently detect and remove masked glitches. Except [1], most of the techniques currently detect or treat each data anomaly in isolation, and they do not exploit patterns of glitches for data cleaning. The detection is also clearly independent and disconnected from the cleaning process. In addition, [14] defined the notion of statistical distortion as an essential metric for measuring the effectiveness of data cleaning strategies since a cleaning method may introduce new errors. Our approach addresses the challenges not addressed by prior work.

VIII. Conclusions

In this paper, we introduced the concept of *masking index*, a statistically rigorous way to quantify the effect of the presence of one type of data glitches on the detection of other types of glitches. We defined two different types of masking, inner and outer masking, to separate out the different effects that one type of glitches has on the detection of another glitch type. Using the fundamental relationship between statistical power and masking, we presented theoretical formulations of the masking index for pairs of different glitch types. In particular, we discussed in detail the effect of missing values on outlier detection.

We illustrated the estimation of the masking index using simulated and real-world data. With simulated data where we can control the occurrence of glitches and establish the ground truth, we can easily estimate the masking index. This allows us to study the behavior of the masking index of a particular method with respect to various characteristics of the glitches. With the real-world data sets, where the ground truth is not known and masking is already present, we proposed a method for estimating the robustness and masking indices of multiple detection methods. This allowed us to identify detection methods that are less affected by masking.

An application of this work is in the area of glitch detection for extremely large data sets where we want to limit the number of detection methods applied to the data. By extracting a smaller subset of the data and performing a similar analysis to this subset, we can identify a small number of detection methods that are less affected by masking to be applied to the large data set.

In the future, we will combine individual masking indices of specified glitch pairs to get an overall masking effect caused by multiple glitch types. We will also propose a generalized linear model based on multiple detection methods to estimate the ground truth. Lastly, we will address the important topic of iterative cleaning, and its effect on the masking index. This will be in the context of statistical distortion introduced by [14]. Essentially, the process of iterative cleaning has to trade-off the reduction in the masking index, along with other data quality criteria, and the statistical distortion caused by the cleaning.

References

[1] L. Berti-Equille, T. Dasu, and D. Srivastava, “Discovery of complex glitch patterns: A novel approach to quantitative data cleaning,” in *ICDE*, 2011, pp. 733–744.

[2] I. Ben-Gal, “Outlier detection,” in *Data Mining and Knowledge Discovery Handbook: A Complete Guide for Practitioners and Researchers*, O. Maimon and L. Rockach, Eds. Kluwer Academic Publishers, 2005.

[3] E. Acuna and C. A. Rodriguez, “Meta analysis study of outlier detection methods in classification,” in *IPSI*, 2004.

[4] H. D., *Identification of Outliers*. Chapman and Hall, 1980.

[5] B. Iglewics and J. Martinez, “Outlier detection using robust measures of scale,” *Journal of Statistical Computation and Simulation*, vol. 15, pp. 285–293, 1982.

[6] L. Davies and U. Gather, “The identification of multiple outliers,” *Journal of the American Statistical Association*, vol. 88, no. 423, pp. 782–792, 1993.

[7] V. Barnett and T. Lewis, *Outliers in Statistical Data*. John Wiley, 1994.

[8] C. R. Rao, *Linear statistical inference and its applications*. John Wiley, 1973.

[9] B. Efron, “Bootstrap methods: Another look at the Jackknife,” *The Annals of Statistics*, vol. 7, pp. 1–26, 1979.

[10] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios, “Duplicate record detection: A survey,” *IEEE Trans. Knowl. Data Eng.*, vol. 19, no. 1, pp. 1–16, 2007.

[11] N. Kushmerick, “Learning to remove internet advertisements,” in *Proceedings of the third annual conference on Autonomous Agents*, ser. AGENTS ’99, 1999, pp. 175–181.

[12] R. Blake and P. Mangiameli, “The effects and interactions of data quality and problem complexity on classification,” *J. Data and Information Quality*, vol. 2, no. 2, pp. 8:1–8:28, Feb. 2011.

[13] H. Xiong, G. Pandey, M. Steinbach, and V. Kumar, “Enhancing data analysis with noise removal,” *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 2, pp. 304–319, 2006.

[14] T. Dasu and J. M. Loh, “Statistical distortion: Consequences of data cleaning,” *PVLDB*, vol. 5, no. 11, pp. 1674–1683, 2012.