Newcomb–Benford law and the detection of frauds in international trade

Andrea Cerioli[a,1], Lucio Barabesi[b], Andrea Cerasa[c], Mario Menegatti[a], and Domenico Perrotta[c,1]

[a]Department of Economics and Management, University of Parma, 43125 Parma, Italy; [b]Department of Economics and Statistics, University of Siena, 53100 Siena, Italy; and [c]European Commission, Joint Research Centre, 21027 Ispra, Italy

Edited by Alex Kossovsky, University of Panama, Panama City, Panama, and accepted by Editorial Board Member Donald B. Rubin October 30, 2018 (received for review April 17, 2018)

The contrast of fraud in international trade is a crucial task of modern economic regulations. We develop statistical tools for the detection of frauds in customs declarations that rely on the Newcomb–Benford law for significant digits. Our first contribution is to show the features, in the context of a European Union market, of the traders for which the law should hold in the absence of fraudulent data manipulation. Our results shed light on a relevant and debated question, since no general known theory can exactly predict validity of the law for genuine empirical data. We also provide approximations to the distribution of test statistics when the Newcomb–Benford law does not hold. These approximations open the door to the development of modified goodness-of-fit procedures with wide applicability and good inferential properties.

statistical antifraud analysis | Newcomb–Benford law | customs fraud | customs valuation | anomaly detection

The contrast of fraud in international trade, and the corresponding protection of national budgets, is a crucial task of modern economic regulations. To give an idea of the volumes involved, in 2016 the customs duties flowing into the European Union (EU) budget amounted to more than 20 billion euros and provided about 15% of the total own resources of the EU. Huge losses thus occur when the value of imported goods is underreported (e.g., ref. 1). Most statistical antifraud techniques for international transactions fall in the class of unsupervised methods, with outlier detection and (robust) cluster analysis playing a prominent role (2–5). The rationale is that the bulk of international trade data are made of legitimate transactions and major frauds may stand out as highly suspicious anomalies. Considerable emphasis is also put on procedures that provide stringent control of the number of false positives (6), since substantial investigations like the one reported in ref. 1 are demanding and time consuming. A related crucial requirement is the ability to deal with massive datasets of traders and to provide—as automatically as possible—a ranking of their degree of anomaly. This information is essential for the design of efficient and effective audit plans, a major task for customs offices.

In this work we consider fraud detection through the Newcomb–Benford law (NBL). This law defines a probability distribution for patterns of significant digits in real positive numbers. It relies on the intriguing fact that in many natural and human phenomena the leading—that is, the first significant—digits are not uniformly scattered, as one could naively expect, but follow a logarithmic-type distribution. We refer to refs. 7–10 for an historical summary of the NBL, an extensive review of its challenging mathematical properties, and a survey of its more relevant applications.

Despite its long history, the mathematical and statistical challenges of the NBL have been recognized only recently. From a mathematical perspective, appropriate versions of the law appear in integer sequences, such as the celebrated Fibonacci sequence (8) or the factorial sequence (11). The law also emerges in the context of floating-point arithmetic (12), while a deep probabilistic study was carried out by Hill (13). A seminal note by Varian (14) suggested the idea that agreement with the NBL could validate the "reasonableness" of data. Since then, it is now rather well known—mainly due to the work of Nigrini (see ref. 7, for a review of such studies)—that the NBL can be used as a forensic accounting and auditing tool for financial data. The law has been shown to be a valuable starting point for forensic accountants and to be applicable in a number of auditing contexts, such as external, internal, and governmental auditing. It has also been found successful for identifying the presence of misconduct in other domains, including the identification of irregularities in electoral data (15, 16), campaign finance (17), and economic data (18).

Although the cited advances may suggest applicability of the NBL to international trade, there remain major unanswered questions that we address in our work. The first one concerns the trustworthiness of the NBL for genuine—that is, nonfraudulent—transactions. As shown in ref. 19, no general known theory can exactly predict whether the NBL should hold in any specific application, whose data-generating process cannot be known with certainty, even in the absence of fraud or

Significance

The detection of frauds is one of the most prominent applications of the Newcomb–Benford law for significant digits. However, no general theory can exactly anticipate whether this law provides a valid model for genuine, that is, nonfraudulent, empirical observations, whose generating process cannot be known with certainty. Our first aim is then to establish conditions for the validity of the Newcomb–Benford law in the field of international trade data, where frauds typically involve huge amounts of money and constitute a major threat for national budgets. We also provide approximations to the distribution of test statistics when the Newcomb–Benford law does not hold, thus opening the door to the development of statistical procedures with good inferential properties and wide applicability.

Author contributions: A. Cerioli, L.B., A. Cerasa, and D.P. designed research; M.M. contributed the study of economic implications of research; A. Cerioli, L.B., A. Cerasa, and D.P. performed research; A. Cerioli, L.B., A. Cerasa, and D.P. contributed new analytic tools; A. Cerioli, L.B., A. Cerasa, and D.P. analyzed data; and A. Cerioli, L.B., M.M., and D.P. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission. A.K. is a guest editor invited by the Editorial Board.

Data deposition: The available pseudo-data files have been deposited at the Athena repository maintained by the Joint Research Centre (JRC). *SI Appendix*, section 3 provides details on how to access them.

See Commentary on page 11.

[1]To whom correspondence may be addressed. Email: andrea.cerioli@unipr.it or domenico.perrotta@ec.europa.eu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1806617115/-/DCSupplemental.

Published online December 10, 2018.

other data manipulations; see also refs. 20–22 for related concerns. Our first goal is then to provide insight on the suitability of the NBL for modeling the distribution of digits of genuine transaction values arising in international trade. We use the Italian import market as a specimen for our study, but our approach is general and can be replicated for any country for which detailed customs data are available. Knowledge of the conditions under which the NBL should be expected to hold in the absence of data manipulation is an essential ingredient for the implementation of large-scale monitoring processes in which tens (or even hundreds) of thousands of traders are screened in an automatic and fast way with the aim of identifying the most suspicious cases. In *SI Appendix*, section 7 we describe a web application that has been developed to assist customs officers and auditors in this screening task, which can be executed in full autonomy on their own datasets. It may instead be very difficult to ascertain whether anomaly should be attributed to fraud or to model failure if the NBL does not provide a suitable model for genuine transactions; see also ref. 23, p. 193, for a similar concern.

Our second goal is to deepen our knowledge of the empirical behavior of NBL-conformance tests by investigating their power under different contamination schemes. The adoption of such tests for antifraud screening is based on the assumption that fabrication of data closely following the law is difficult and that fraudsters might be biased toward simpler digit distributions, such as the discrete uniform or the Dirac. We also quantify the corresponding false positive rates, to make explicit the different and possibly conflicting facets that empirical researchers have to balance in practice.

The third aim of our work is to provide corrections to test statistics when the NBL does not hold. This is typically the case for traders who operate on a limited number of products, so that there is not enough variability in their transactions. Even if the NBL is not a suitable model for genuine transaction digits, the conformance tests based on our modified statistics have the appropriate empirical size in the absence of data manipulation, while the usual tests turn out to be potentially very liberal. We argue that, having the required size under general trade conditions and being competitive in terms of power, the conformance tests based on our modified statistics are recommended. Therefore, they extend the applicability of large-scale monitoring processes of international trade data to a wider range of practical situations.

The NBL

Statistical Background. Let $D_1(x), D_2(x), \ldots$, be the first, the second, …, significant digit of the positive real number x. Let X be a positive real random variable defined on the probability space $(\Omega, \mathcal{F}, P)$. The NBL implies (and vice versa) that the following joint probability function holds for each $k \in \mathbb{Z}^+$,

$$\rho_k(d_1, \ldots, d_k) = P(D_1(X) = d_1, \ldots, D_k(X) = d_k) = \log_{10}\left(1 + \frac{1}{\sum_{l=1}^{k} 10^{k-l} d_l}\right), \quad [1]$$

where $d_1 \in \{1, \ldots, 9\}$ and $d_l \in \{0, \ldots, 9\}$ for $l = 2, \ldots, k$. A practically important special case is that of the first two significant digits ($k = 2$), for which Eq. **1** reduces to

$$\rho_2(d_1, d_2) = \log_{10}\left(1 + \frac{1}{10d_1 + d_2}\right). \quad [2]$$

Similarly, the marginal probability function of $D_1(X)$ is

$$P(D_1(X) = d_1) = \log_{10}\left(1 + \frac{1}{d_1}\right),$$

while the marginal probability function of $D_2(X)$ is

$$P(D_2(X) = d_2) = \sum_{d_1=1}^{9} \log_{10}\left(1 + \frac{1}{10d_1 + d_2}\right).$$

We refer to ref. 24 for a summary of the mechanisms that give rise to NBL-distributed data in accounting and finance. Among these, there are several statistical motivations for adopting the NBL as a model for the digits appearing in genuine international transactions. A major methodological basis relies on a limit theorem derived by Hill (13), to which we refer for the technical details. A key mathematical concept is that of a random probability measure, which is a function $\mathbb{P} : \Omega \to \mathcal{M}$—where $\mathcal{M}$ is the space of probability measures on $\mathbb{R}$—defined on the underlying probability space $(\Omega, \mathcal{F}, P)$. For each Borel set B the function $\omega \mapsto \mathbb{P}(\omega)(B)$ is a random variable; that is, $\mathbb{P}(\omega)$ is a probability measure on $\mathbb{R}$ for each $\omega \in \Omega$. Another important related concept is that of a sequence of $\mathbb{P}$-random M samples, where $M \in \mathbb{Z}^+$. It is a sequence $(X_n)_{n\geq 1}$ of random variables defined on $(\Omega, \mathcal{F}, P)$ such that, for each $\omega \in \Omega$, the first M random variables are drawn independently from the same random probability distribution $\mathbb{P}_1(\omega)$, selected according to the random probability measure $\mathbb{P}$, the M subsequent random variables are drawn independently from the same random probability distribution $\mathbb{P}_2(\omega)$, in turn selected according to the random probability measure $\mathbb{P}$, and so on. Hill's limit theorem then states that, if $\mathbb{P}$ satisfies some invariance conditions related to either the scale or the base of measurement, for each $M \in \mathbb{Z}^+$ the $\mathbb{P}$-random M-samples sequence $(X_n)_{n\geq 1}$ converges to the NBL with probability one. That is, for each $k \in \mathbb{Z}^+$ and for $i = 1, \ldots, n$, as $n \to \infty$

$$\frac{\text{card}\{i : D_1(X_i) = d_1, \ldots, D_k(X_i) = d_k\}}{n} \overset{a.s.}{\to} \rho_k(d_1, \ldots, d_k). \quad [3]$$

A second reason for adopting the NBL is that multiplicative processes—which are at the heart of many financial data—generate NBL-distributed data. More precisely, if $(X_n)_{n\geq 1}$ is a sequence of independent and identically distributed random variables such that $P(X_1 = 0) = 0$, as $n \to \infty$ the sequence $(\prod_{i=1}^{n} X_i)_{n\geq 1}$ converges to the NBL with probability one (theorem 8.16 in ref. 8). It can be shown that convergence is extremely fast since it is exponential in n (25). It is also remarkable that, given two independent random variables X and Y only one of which follows the NBL, the product XY is distributed according the NBL provided that $P(XY = 0) = 0$ (theorem 8.12 in ref. 8). Finally, NBL-distributed data may also originate from random variables raised to integer powers. If X is an absolutely continuous random variable, as $n \to \infty$ the sequence $(X^n)_{n\geq 1}$ converges to the NBL with probability one (theorem 8.8 in ref. 8).

Relevance for International Trade. Our applied focus is on transactions involving EU traders; we refer to *SI Appendix*, sections 3 and 7 for the institutional regulations supporting their analysis. By international trade data we mean the data collected by EU member states for imports and exports that are declared by national traders and shipping agents using the form called the Single Administrative Document (SAD). The value that we analyze for antifraud purposes is the "statistical value" reported in each SAD, which also includes the costs of insurance and freight (CIF) and is given in euros by taking into account the exchange rate (26). Our interest is then on random variables $X_1, \ldots, X_n$ defined on the product space

$$X_i = U_i Q_i, \qquad i = 1, \ldots, n, \quad [4]$$

where U_i and Q_i are nonnegative random variables representing the (CIF-type) unit price in euros and the traded quantity in transaction i. If we rephrase **[3]** in the context of trade,

n corresponds to the number of transactions made by the trader of interest, so that $X_1, \ldots, X_n$ is the available sample of transaction values, and the ratio $m = n/M$ is the corresponding number of traded goods (provided that m is an integer).

There are different economic reasons suggesting that the distribution of the significant digits contained in $X_1, \ldots, X_n$ may, under some conditions, be well approximated by the NBL. First, markets are hit by specific shocks and show peculiar reactions to common shocks (27). This, coupled with differences in the trader size and product quality, generates different economic processes for prices and quantities determination, which imply in turn that the observed data of prices and quantities may be described by different trader-specific probability distributions, not exactly predictable in advance. In view of **[3]**, it is then sensible to anticipate good conformance to the NBL when a trader operates by importing or exporting a sufficiently large number of different goods, even if none of the product-specific marginal distributions of digits follows the law. The economic literature also shows that traders have different degrees of market power. Trading operations are affected by market and country features, such as different trade costs and different access to credit (e.g., ref. 28). Therefore, transactions made with different counterparties may be characterized by different economic processes, yielding distributions for transaction values that can be conceived to vary randomly from one product to another for each trader. The significant-digit distribution in international transactions can thus be expected to adhere to the NBL when the trader makes a sufficiently large number of operations, with a sufficiently large number of counterparties, possibly located in different countries.

A Contamination Model for Fraud

The Model. We phrase our antifraud approach within the framework of a trader-specific contamination model where each fraud corresponds to an outlier. For this purpose, we need a slight change in notation and we write n_t for the number of transactions made by trader t, which operates on m_t distinct products and for which the positive random variable $X^{(t)}$ now represents a transaction value. We then define

$$\pi_k^{(t)}(d_1, \ldots, d_k) = P(D_1(X^{(t)}) = d_1, \ldots, D_k(X^{(t)}) = d_k),$$

and let T denote the total number of traders in the market.

For $t = 1, \ldots, T$ and each $k \in \mathbb{Z}^+$, the general form of our contamination model is

$$\pi_k^{(t)}(d_1, \ldots, d_k) = (1 - \tau_t)\Psi_k^{(t)}(d_1, \ldots, d_k) + \tau_t \Upsilon_k^{(t)}(d_1, \ldots, d_k), \quad \textbf{[5]}$$

where $\Psi_k^{(t)}(d_1, \ldots, d_k)$ is the probability of observing $\{D_1(X^{(t)}) = d_1, \ldots, D_k(X^{(t)}) = d_k\}$ in the absence of fraud, $\Upsilon_k^{(t)}(d_1, \ldots, d_k)$ is the probability of the same event for a manipulated transaction, and $0 \le \tau_t \le 1$ is the probability of fraud for trader t. Although it is convenient to work in the digit space through $\pi_k^{(t)}(d_1, \ldots, d_k)$, model **5** has a counterpart in the transaction space defined by $X^{(t)}$. The latter is given in *SI Appendix*, section 1.

Model **5** provides a principled framework for antifraud analysis of international trade data. Indeed, trader t may be considered a potential fraudster if the null hypothesis

$$H_0^{(t)} : \tau_t = 0 \quad \textbf{[6]}$$

is rejected, in favor of the alternative $H_1^{(t)} : \tau_t > 0$, based on n_t independent copies of $X^{(t)}$, say $X_1^{(t)}, \ldots, X_{n_t}^{(t)}$.

A useful tractable version of contamination model **5** assumes that the probability of observing a given k-ple of digits in a genuine transaction of trader t depends on the trader features only through the values of m_t and n_t; that is,

$$\Psi_k^{(t)}(d_1, \ldots, d_k) := \Psi_k^{(m_t, n_t)}(d_1, \ldots, d_k).$$

Therefore, for each $k \in \mathbb{Z}^+$, the model becomes

$$\pi_k^{(t)}(d_1, \ldots, d_k) = (1 - \tau_t)\Psi_k^{(m_t, n_t)}(d_1, \ldots, d_k) + \tau_t \Upsilon_k^{(t)}(d_1, \ldots, d_k), \quad \textbf{[7]}$$

with **[6]** again stating the absence of fraud. Model **7** implies that the random vector $(D_1(X^{(t)}), \ldots, D_k(X^{(t)}))$ is independent of any other trader-specific random variable, given the values of m_t and n_t. Although this structure is clearly an approximation, it is coherent with the discussion about the economic elements that make the NBL a plausible model for the digit distribution in genuine international transactions.

A further bonus of models **5** and **7** is that they make clear the antifraud advantages of our methodology over the often uninformative analysis of aggregated data, as given, for example, in ref. 18. In the latter instance, for each $k \in \mathbb{Z}^+$, the underlying contamination model would be

$$\pi_k(d_1, \ldots, d_k) = (1 - \tau)\Psi_k(d_1, \ldots, d_k) + \tau \Upsilon_k(d_1, \ldots, d_k),$$

where the quantities involved are now constant for the whole (product-specific) market. Testing the hypothesis that $\tau = 0$ in this restricted model requires a sample $X_1, \ldots, X_T$ obtained from T traders, for which just one replicate is available. However, the inferential conclusion that $\tau > 0$ is much less informative than rejection of **[6]** for some $t \in \{1, \ldots, T\}$. In fact, $\tau > 0$ yields no information on the specific traders that are responsible for rejection and identification of the fraudsters must be left to further nonstatistical investigations. Another notable advantage is that models **5** and **7** acknowledge the existence of a trader-specific propensity to fraud.

Testing the Absence of Fraud. The usual hypothesis of interest in the antifraud literature (7, 10) is

$$H_0^{(t)} : \pi_k^{(t)}(d_1, \ldots, d_k) = \rho_k(d_1, \ldots, d_k), \qquad \forall\, k \in \mathbb{Z}^+, \quad \textbf{[8]}$$

which corresponds to **[6]** when $\Psi_k^{(t)}(d_1, \ldots, d_k)$ is the NBL. Several statistics exist for testing **[8]** for a given value of k, the simplest one being the χ^2 statistic

$$V_{\{1,\ldots,k\}}^{(t)} = \sum_{d_1, \ldots, d_k} \frac{\left(N_k^{(t)}(d_1, \ldots, d_k) - n_t \rho_k(d_1, \ldots, d_k)\right)^2}{n_t \rho_k(d_1, \ldots, d_k)}, \quad \textbf{[9]}$$

where $N_k^{(t)}(d_1, \ldots, d_k)$ is the frequency of the k-ple $(d_1, \ldots, d_k)$ in the sample of n_t transactions for trader t. It is a standard result that, as $n_t \to \infty$, $V_{\{1,\ldots,k\}}^{(t)} \xrightarrow{\mathcal{L}} \chi^2_\nu$ when **[8]** is true, with $\nu = 9 \times 10^{k-1} - 1$. In practice only NBL marginals of low order are analyzed. The two-digit version of **[9]**, that is, $V_{\{1,2\}}^{(t)}$, tests the fit to the 2D marginal of the NBL given in **[2]**, while the corresponding 1D marginal hypotheses are tested through $V_{\{1\}}^{(t)}$ and $V_{\{2\}}^{(t)}$, respectively.

In our empirical study we also consider the multiple-stage approach proposed by Barabesi et al. (6) with the aim of introducing a more stringent control on the proportion of false discoveries. This approach tests a decreasing sequence of lower-dimensional marginals of the NBL through their exact

conditional distributions. Specifically, in the simple two-step version that we consider here, the method of Barabesi et al. (6) first tests the two-digit marginal **2** of the NBL by comparing $V^{(t)}_{\{1,2\}}$ to the quantiles of its exact distribution under the null, which are approximated through an efficient Monte Carlo scheme. Then, if the 2D NBL is rejected, the fit to the 1D marginals is tested by $V^{(t)}_{\{1\}}$ and $V^{(t)}_{\{2\}}$. These lower-dimensional tests use the exact conditional distributions of $V^{(t)}_{\{1\}}$ and $V^{(t)}_{\{2\}}$, given rejection of the 2D hypothesis, instead of their marginal ones. Type-I error rates are thus controlled at the prescribed level (e.g., 1%) at each step of the procedure, both in the two-digit and in the one-digit tests. Furthermore, the outcome on the one-digit tests reveals which digit is responsible for nonconformance to **[2]**.

Since χ^2 tests may also have some shortcomings (ref. 10, chap. 37), additional procedures not based on **[9]** and less formal methods are considered in *SI Appendix*, sections 5 and 6. Qualitative findings are similar in all cases. Nevertheless, for our purposes it is instructive to look at the results for χ^2 tests, because their distribution (either exact or asymptotic) is known under the NBL. We can thus look at the agreement between the empirical and the nominal distribution of the test statistics to assess whether genuine transactions actually follow the law, that is, if $\Psi^{(t)}_k(d_1,\dots,d_k)$ in **[5]** (or $\Psi^{(m_t,n_t)}_k(d_1,\dots,d_k)$ in **[7]**) is the NBL.

Adequacy of the NBL for Trade Data

Although the theoretical results sketched in the statistical background and the subsequent economic arguments broadly motivate the adoption of the NBL as a sensible model for genuine transactions in the context of international trade, it is unclear how they may fit to empirical transactions whose generating mechanism cannot be exactly known and obviously involves only a finite number of terms. One goal of our study is then to provide evidence on the quality of the NBL assumption **1** to the digit distribution of transaction values for noncheating traders that operate in real international markets. For this purpose, we assume that our contamination model holds with $\tau_t = 0$ for each trader. We also take **[7]** as a sensible and practically workable approximation to this model in the absence of a priori information on the trader.

We simulate nonmanipulated statistical values, according to definition **4**, for $T^\dagger$ "idealized" traders in each relevant configuration of trade represented by a pair (m_t, n_t). For this aim, we sample transactions with replacement from the Cartesian product spaces

$$\mathcal{X}_j = \mathcal{U}_j \times \mathcal{Q}_j, \qquad j = 1,\dots,G, \qquad \textbf{[10]}$$

where $\mathcal{U}_j = \{u_1,\dots,u_{n_j}\}$ and $\mathcal{Q}_j = \{q_1,\dots,q_{n_j}\}$ denote the sets of unit prices (CIF-type) and traded quantities, respectively, originated in all of the market transactions involving good j, n_j is the number of such transactions, and G is the total number of goods in the market. The details of the simulation algorithm are reported in *SI Appendix*, section 2. In our experimental setting the values of m_t and n_t are fixed by design, while in empirical analysis we instead condition on the observed values of m_t and n_t for the trader under scrutiny. We replicate genuine international trading behavior in one specific EU market by picking unit price and traded quantity at random from the database of one calendar year Italian customs declarations, after appropriate trader and product anonymization making it impossible to infer the features of individual operators. Two databases of simulated transactions (pseudo-datasets) similar to those analyzed in this work can be accessed through *SI Appendix*, section 3, where their structure is explained. A description of our code is also given in *SI Appendix*, section 3.

For each idealized trader t and a chosen value of k, we compare the observed distribution of digits to the theoretical NBL values **1** through the test statistic $V^{(t)}_{\{1,\dots,k\}}$. This statistic will be asymptotically distributed as χ^2_ν if $\Psi^{(m_t,n_t)}_k(d_1,\dots,d_k)$ is indeed the NBL. Furthermore, its exact distribution under the k-digit NBL hypothesis can be approximated to an arbitrary degree of accuracy through the Monte Carlo approach of Barabesi et al. (6). We thus take the discrepancy between the estimated distribution of $V^{(t)}_{\{1,\dots,k\}}$, computed by averaging over the $T^\dagger$ Monte Carlo replicates of t, and its reference null distribution, say $F_{V^{(t)}_{\{1,\dots,k\}}}$, as a measure of the adequacy of the NBL assumption in model **7**. Formally, let ζ_γ be the γ quantile of $F_{V^{(t)}_{\{1,\dots,k\}}}$ and let I_C denote the indicator function of a given set C. Our Monte Carlo estimate is computed as

$$\hat{\alpha} = \frac{1}{T^\dagger}\sum_{t=1}^{T^\dagger} I_{]\zeta_{1-\alpha},+\infty[}\left(V^{(t)}_{\{1,\dots,k\}}\right), \qquad \textbf{[11]}$$

for α in the usual range of significance levels. Although a value of $\hat{\alpha}$ close to α does not imply that the empirical distribution of $V^{(t)}_{\{1,\dots,k\}}$ is well approximated by $F_{V^{(t)}_{\{1,\dots,k\}}}$ over all its support, it tells us that the approximation is satisfactory for the purpose for which $V^{(t)}_{\{1,\dots,k\}}$ is computed in antifraud analysis. The insight that we gain from our study is twofold. First, we shed light on the trading configurations—represented in terms of pairs (m_t, n_t)—that ensure close agreement between $\Psi^{(m_t,n_t)}_k(d_1,\dots,d_k)$ and the NBL in the market from which all of the sets $\mathcal{U}_j$ and $\mathcal{Q}_j$ are obtained. Second, we explore the effect of sparseness of digit counts on the distribution of $V^{(t)}_{\{1,\dots,k\}}$ when n_t is small or moderate.

The bulk of our results deal with the simple first-digit statistic $V^{(t)}_{\{1\}}$, which is likely to be method of choice by many antifraud practitioners in automated large-scale auditing processes. As a reference, we also provide the estimated test sizes for the two-stage (TS) version of the procedure of Barabesi et al. (6) and for the two-digit statistic $V^{(t)}_{\{1,2\}}$. The former is intended to be a reasonable compromise between simplicity of use and strong reduction in the rate of false detections, while the latter is often recommended in applications with not-too-small sample sizes (ref. 7, p. 79). We estimate test sizes using **[11]** for a wide range of pairs (m_t, n_t), with $m_t \le n_t$. The chosen grid represents the features of some of the most relevant traders in the empirical analysis of customs declarations. In fact, the importers for which $n_t < 50$ cover less than 14% of the recorded transactions in our customs database and an even smaller quota in terms of traded values. Very big traders are not common: To give an idea, $n_t > 2{,}000$ for less than 0.1% of the importers in the database, and almost 40% of the recorded transactions refer to traders with $50 \le n_t \le 2{,}000$. We present only the findings for the case $\alpha = 0.01$, similar conclusions being valid for other significance levels.

Table 1 displays the estimated sizes of the test of the first-digit marginal hypothesis for both $V^{(t)}_{\{1\}}$ (using the quantiles of its asymptotic distribution) and TS. These estimates are computed on $T^\dagger = 85{,}500$ idealized noncheating traders, pooled across different scenarios with the same pair (m_t, n_t). One striking feature of the reported values of $\hat{\alpha}$ in Table 1 is that they vary considerably according to the specific trading configuration. This result clearly supports the conjecture that in a realistic market scenario both m_t and n_t are crucial factors in determining the adequacy of the NBL as a valid model for the empirical digit distribution in the absence of data manipulation. If only one digit is considered, a sample size of $n_t = 50$ transactions can be considered

Table 1. Estimated test sizes (Eq. 11) for the first-digit statistic $V^{(t)}_{\{1\}}$, using the asymptotic quantile $\chi^2_{8,0.99}$, and for the TS version of the procedure of Barabesi et al. (6), based on $T^\dagger = 85,500$ Monte Carlo replicates for each configuration (m_t, n_t), with $m_t \le n_t$

No. of transactions	Test	m_t								
		1	5	10	20	40	80	100	200	500
$n_t = 50$	$V^{(t)}_{\{1\}}$	0.053	0.027	0.018	0.014	0.011	—	—	—	—
	TS	0.024	0.003	0.001	0.000	0.000	—	—	—	—
$n_t = 100$	$V^{(t)}_{\{1\}}$	0.071	0.045	0.027	0.016	0.012	0.011	0.011	—	—
	TS	0.049	0.013	0.004	0.001	0.000	0.000	0.000	—	—
$n_t = 200$	$V^{(t)}_{\{1\}}$	0.094	0.069	0.047	0.026	0.016	0.012	0.011	0.010	—
	TS	0.070	0.035	0.013	0.003	0.001	0.000	0.000	0.000	—
$n_t = 500$	$V^{(t)}_{\{1\}}$	0.132	0.126	0.097	0.062	0.031	0.017	0.016	0.012	0.010
	TS	0.103	0.084	0.049	0.017	0.003	0.000	0.000	0.000	0.000

Model **7** holds with $\tau_t = 0$ for each trader. The nominal test size is $\alpha = 0.01$.

sufficiently large to justify the asymptotic χ^2_8 approximation to the distribution of $V^{(t)}_{\{1\}}$ and the adoption of the NBL as a reasonable model for $\Psi_1^{(m_t,n_t)}(d_1)$, provided that the number of traded products is large as well (around 20, say). Similar findings hold for all of the pairs (m_t, n_t) taken into account in our experiment and provide an empirical verification of the speed of convergence to the NBL anticipated by the asymptotic framework of Hill's result **3**. An interesting remark is that $\hat{\alpha}$ for $V^{(t)}_{\{1\}}$ is closer to α when $m_t = n_t$, thus suggesting that convergence in **[3]** is faster when $M = 1$. On the other hand, TS yields a very conservative test when the NBL provides a satisfactory model. This result is hardly surprising, since TS tests the first-digit hypothesis at nominal size α in the conditional distribution of $V^{(t)}_{\{1\}}$, given previous rejection of the two-digit NBL hypothesis. In *SI Appendix*, section 5, we also investigate the fit of the whole empirical distribution of $V^{(t)}_{\{1\}}$ to the nominal χ^2_8 distribution.

Our results point to the conclusion that the NBL is not a satisfactory model when m_t is much smaller than n_t. This statement is verified consistently over all market configurations and does not depend on the specific testing methodology. Indeed, also the potentially very conservative TS procedure can become considerably liberal if $m_t \ll n_t$. The same is true for other adjustments to $V^{(t)}_{\{1\}}$ that control for multiplicity of tests among traders, not reported here. We argue that lack of variability in the transactions made by trader t is the main reason for the discrepancy between the NBL and $\Psi_k^{(m_t,n_t)}(d_1, \ldots, d_k)$ when m_t is small. Whatever the interpretation, our simulation results confirm that the asymptotic framework set by **[3]** does not hold if $m_t = o(n_t)$, requiring instead $m_t = O(n_t)$. Our results also quantify how much deleterious can be the effect of keeping m_t fixed on the distribution of test statistics. Indeed, they show that in this setting an increase of the sample size n_t worsens the situation, since it points to a "wrong" asymptotic direction. The clear message is then that standard conformance tests, such as $V^{(t)}_{\{1,\ldots,k\}}$, should not be used for antifraud purposes when $m_t \ll n_t$, because the hypotheses **6** and **8** cannot be taken any longer to be equivalent.

We conclude this section with a glimpse of the performance of the two-digit statistic $V^{(t)}_{\{1,2\}}$, when either the asymptotic quantile $\chi^2_{89,0.99}$ or the exact 0.99 quantile from Barabesi et al. (6) is used. The estimated test sizes, now based on $T^\dagger = 28,500$ Monte Carlo replicates for each configuration (m_t, n_t), are reported in Table 2. As expected, convergence to the χ^2_{89} distribution is slower than convergence to χ^2_8 in the one-digit case. The adoption of exact quantiles should thus be preferred with $V^{(t)}_{\{1,2\}}$, except in the instance of large values of both n_t and m_t. Our results confirm the relationship between accuracy of the NBL approximation and the ratio m_t/n_t, suggesting $m_t \ge 0.2n_t$ as a sensible rule of thumb when the exact quantiles are used. They also provide a clue of the strategy to be adopted with more complex large-k procedures.

Enemy Brothers: Power and False Positive Rate

When model **7** holds with $\tau_t > 0$ for one or more traders, we write $\mathcal{T}_{\mathrm{NF}} = \{t : \tau_t = 0\}$ and $\mathcal{T}_{\mathrm{F}} = \{t : \tau_t > 0\}$ for the sets corresponding to noncheating traders and fraudsters, respectively.

Table 2. Estimated test sizes (Eq. 11) for the two-digit statistic $V^{(t)}_{\{1,2\}}$, using the asymptotic quantile $\chi^2_{89,0.99}$ (As) and the exact 0.99 quantile (Ex) from Barabesi et al. (6), based on $T^\dagger = 28,500$ Monte Carlo replicates for each configuration (m_t, n_t), with $m_t \le n_t$

No. of transactions	Test	m_t								
		1	5	10	20	40	80	100	200	500
$n_t = 50$	As	0.064	0.039	0.035	0.029	0.026	—	—	—	—
	Ex	0.040	0.017	0.013	0.011	0.010	—	—	—	—
$n_t = 100$	As	0.083	0.048	0.033	0.023	0.021	0.020	0.019	—	—
	Ex	0.068	0.032	0.019	0.013	0.011	0.010	0.010	—	—
$n_t = 200$	As	0.102	0.069	0.043	0.025	0.018	0.014	0.016	0.014	—
	Ex	0.095	0.059	0.034	0.018	0.012	0.010	0.011	0.009	—
$n_t = 500$	As	0.141	0.125	0.087	0.052	0.027	0.016	0.014	0.012	0.012
	Ex	0.137	0.120	0.082	0.047	0.023	0.013	0.012	0.010	0.009

Model **7** holds with $\tau_t = 0$ for each trader. The nominal test size is $\alpha = 0.01$.

Table 3. Uniform contamination model 12

Trade configuration	Test	$\varsigma = 0.05$						$\varsigma = 0.10$					
		$\tau_t = 0.2$		$\tau_t = 0.5$		$\tau_t = 0.8$		$\tau_t = 0.2$		$\tau_t = 0.5$		$\tau_t = 0.8$	
		P	FPR	P	FPR	P	FPR	P	FPR	P	FPR	P	FPR
$n_t = 50$	$V^{(t)}_{\{1\}}$	0.034	0.865	0.196	0.546	0.586	0.302	0.030	0.779	0.200	0.346	0.574	0.178
$m_t = 50$	TS	0.002	0.000	0.008	0.000	0.154	0.013	0.000	1	0.019	0.000	0.133	0.007
$n_t = 100$	$V^{(t)}_{\{1\}}$	0.058	0.788	0.436	0.297	0.938	0.184	0.043	0.705	0.425	0.175	0.924	0.097
$m_t = 100$	TS	0.004	0.000	0.070	0.054	0.574	0.003	0.002	0.667	0.063	0.000	0.539	0.002
$n_t = 200$	$V^{(t)}_{\{1\}}$	0.060	0.778	0.810	0.179	1	0.151	0.097	0.484	0.801	0.109	1	0.097
$m_t = 200$	TS	0.006	0.500	0.356	0.000	0.964	0.002	0.005	0.444	0.345	0.003	0.959	0.004
$n_t = 500$	$V^{(t)}_{\{1\}}$	0.272	0.401	1	0.160	1	0.154	0.281	0.226	1	0.069	1	0.081
$m_t = 500$	TS	0.028	0.263	0.932	0.000	1	0.004	0.029	0.065	0.928	0.000	1	0.000

Shown are estimated power (P) and false positive rate (FPR) for the first-digit statistic $V^{(t)}_{\{1\}}$, using the asymptotic quantile $\chi^2_{8,0.99}$, and for the TS version of the procedure of Barabesi et al. (6), based on $T^\dagger = 10,000$ Monte Carlo replicates for each pair (m_t, n_t). The nominal test size is $\alpha = 0.01$.

Power (P) is defined as the proportion of traders in $\mathcal{T}_{\mathrm{F}}$ that are correctly identified as potential fraudsters. The false positive rate (FPR) is the proportion of rejections of the null hypothesis **6** that turn out to be wrong, since they refer to traders that belong to $\mathcal{T}_{\mathrm{NF}}$. Both performance measures play a crucial role when antifraud analysis is put into practice. In our simulations we compare the results under different contaminant distributions $\Upsilon^{(t)}_k(d_1, \ldots, d_k)$, with $k = 2$.

Our first contamination instance assumes that the first two digits of $\tau_t n_t$ transactions from trader $t \in \mathcal{T}_{\mathrm{F}}$ are generated according to the discrete uniform distribution on $\{10, \ldots, 99\}$. Therefore,

$$\pi^{(t)}_2(d_1, d_2) = (1 - \tau_t)\Psi^{(m_t, n_t)}_2(d_1, d_2) + \tau_t \frac{1}{90}, \qquad [12]$$

for $d_1 \in \{1, \ldots, 9\}$ and $d_2 \in \{0, \ldots, 9\}$. The uniform distribution provides an unfavorable scenario for fraud detection, since $\Upsilon^{(t)}_2(d_1, d_2)$ is then close to the NBL marginal probability **2** for most digit pairs (d_1, d_2). Our second contamination scheme instead concentrates frauds on a specific digit pair, say $(\bar{d}_1, \bar{d}_2)$, randomly selected from the discrete uniform distribution on $\{10, \ldots, 99\}$. The contaminated model thus becomes

$$\pi^{(t)}_k(d_1, d_2) = (1 - \tau_t)\Psi^{(m_t, n_t)}_2(d_1, d_2) + \tau_t I_{\{\bar{d}_1, \bar{d}_2\}}(d_1, d_2). \qquad [13]$$

Although this Dirac-type contamination may at first sight appear extreme, our experience with manipulated declarations is that similar patterns may arise rather frequently among the transactions found to be fraudulent, especially when contamination is due to the attempt to circumvent threshold-depending duties, either "ad valorem"—that is, computed as a percentage of the declared value—or fixed. In fact, the attempt to declare quantities below the threshold (or above it, according to the specific regulation) typically produces a bias in the corresponding values similar to that represented by a Dirac-type model. Other instances of contamination are considered in *SI Appendix*, section 4.

We consider the simplified case where τ_t is the same for each $t \in \mathcal{T}_{\mathrm{F}}$. We take $\tau_t = 0.2, 0.5, 0.8$, to represent three increasing levels of individual propensity to fraud. We also define the proportion of fraudsters in the whole market as

$$\varsigma = \frac{\mathrm{card}(\mathcal{T}_{\mathrm{F}})}{\mathrm{card}(\mathcal{T})},$$

where $\mathcal{T} = \mathcal{T}_{\mathrm{NF}} \bigcup \mathcal{T}_{\mathrm{F}}$ is the set of all traders. We fix $\varsigma = 0.05, 0.1$, to investigate the effect of different degrees of fraud diffusion in the market. Our estimates of P and FPR are based on $T^\dagger = 10,000$ idealized traders, independently generated in each configuration. Nonmanipulated transactions are again simulated with the algorithm described in *SI Appendix*, section 2. We restrict our analysis to the market configurations for which the NBL approximation to $\Psi^{(m_t, n_t)}_2(d_1, d_2)$ is good, and the empirical test sizes closely match the nominal one, to avoid confounding between power and lack of fit. We give results only for the configurations with $m_t = n_t$. Pairs where m_t is of the same order of magnitude as n_t yield qualitatively similar findings and are not reported.

Table 3 shows the estimated values of P and FPR under the uniform contamination model **12** for $V^{(t)}_{\{1\}}$, using the asymptotic quantile $\chi^2_{8,0.99}$, and for the TS version of the procedure of Barabesi et al. (6). Not surprisingly, the detection rates are low in the case of sporadic contamination ($\tau_t = 0.2$). It is apparent that no statistical method can be expected to have high power against "well-masked" frauds, unless the number of contaminated transactions becomes relatively large. Indeed, it is clearly seen that P rapidly grows with both τ_t and n_t, leading to almost sure detection of fraudsters even through the potentially very conservative TS procedure (e.g., when $\tau_t = 0.8$ and $n_t \geq 200$). Both methods thus prove to be able to identify the traders belonging to $\mathcal{T}_{\mathrm{F}}$ if there is enough information on the contaminant distribution in the available data, also in the unfavorable framework provided by **[12]**. The value of FPR is much higher with $V^{(t)}_{\{1\}}$, as expected, except in some instances of low contamination, where the number of hypotheses **6** rejected by TS is very small and the estimate of FPR is overwhelmed by its sampling variability. The choice between $V^{(t)}_{\{1\}}$ and TS should then depend on the user's attitude toward FPR and toward the power reduction implied by TS in situations of intermediate contamination. The value of ς does not have a major impact on P, thus suggesting that our procedures can be equally effective in detecting isolated fraudsters and more diffuse illegal trading behavior. However, a considerable increase in FPR is to be expected in the former situation, especially for $V^{(t)}_{\{1\}}$.

Table 4 repeats the analysis under the Dirac-type scheme **13**. The contaminant distribution is now well separated from $\Psi^{(m_t, n_t)}_2(d_1, d_2)$ and both methods generally have excellent detection properties, with some minor differences only in the problematic case $\tau_t = 0.2$. However, FPR is much higher for $V^{(t)}_{\{1\}}$. In such contamination frameworks the TS procedure thus comes closer to performing like an "ideal" test, leading to the

Table 4. The same as Table 3, but now for contamination model 13

Trade configuration	Test	$\varsigma = 0.05$						$\varsigma = 0.10$					
		$\tau_t = 0.2$		$\tau_t = 0.5$		$\tau_t = 0.8$		$\tau_t = 0.2$		$\tau_t = 0.5$		$\tau_t = 0.8$	
		P	FPR	P	FPR	P	FPR	P	FPR	P	FPR	P	FPR
$n_t = 50$	$V^{(t)}_{\{1\}}$	0.712	0.218	0.998	0.199	1	0.184	0.696	0.121	0.996	0.092	1	0.108
$m_t = 50$	TS	0.520	0.763	1	0.002	1	0.002	0.555	0.005	1	0.003	1	0.001
$n_t = 100$	$V^{(t)}_{\{1\}}$	0.876	0.189	1	0.188	1	0.145	0.891	0.095	1	0.083	1	0.081
$m_t = 100$	TS	0.972	0.008	1	0.004	1	0.000	0.980	0.001	1	0.003	1	0.001
$n_t = 200$	$V^{(t)}_{\{1\}}$	0.972	0.169	1	0.167	1	0.150	0.967	0.091	1	0.079	1	0.076
$m_t = 200$	TS	1	0.004	1	0.006	1	0.000	1	0.002	1	0.002	1	0.000
$n_t = 500$	$V^{(t)}_{\{1\}}$	1	0.171	1	0.158	1	0.176	1	0.078	1	0.094	1	0.071
$m_t = 500$	TS	1	0.006	1	0.000	1	0.000	1	0.002	1	0.003	1	0.003

identification of most potential fraudsters with a very small number of false alarms. The effect of ς is still minor on P, while it is more noticeable on FPR for $V^{(t)}_{\{1\}}$.

Corrections to Goodness-of-Fit Statistics

We now focus on the trading configurations for which the NBL does not provide a satisfactory representation of the genuine digit distribution $\Psi_k^{(m_t,n_t)}(d_1,\ldots,d_k)$, that is, when $m_t \ll n_t$. In this case, the reported distributional results are no longer valid for $V^{(t)}_{\{1,\ldots,k\}}$ or for the exact Monte Carlo approach of Barabesi et al. (6). The true probability $\Psi_k^{(m_t,n_t)}(d_1,\ldots,d_k)$ should replace the NBL version of $\pi_k^{(t)}(d_1,\ldots,d_k)$ in **[9]** to obtain valid tests of hypothesis **6**. Since $\Psi_k^{(m_t,n_t)}(d_1,\ldots,d_k)$ is unknown, we resort to our Monte Carlo algorithm for simulating nonfraudulent transactions and we compute a model-free approximation to the null distribution function of $V^{(t)}_{\{1,\ldots,k\}}$. This approximation is then used to obtain a test of **[6]**. Similar testing procedures have proved to be useful in other domains, in the case of correlated observations and other distributional misspecifications (e.g., ref. 29 and the references therein).

If t is the trader of interest, let t^* be an idealized noncheating trader such that $t^* \neq t$, while $m_{t^*} = m_t$ and $n_{t^*} = n_t$. The set of transactions for trader t^* is randomly generated according to the algorithm described in *SI Appendix*, section 2, and the resulting statistical values are collected in vector $x^{(t^*)}$, say. Correspondingly, let $V^{(t^*)}_{\{1,\ldots,k\}}$ be the test statistic **9** computed for trader t^*. Under model **7**, the significant-digit random variables associated to the elements of $x^{(t^*)}$ can be considered as independent copies of those associated to the elements of $X^{(t)}$, in the absence of data manipulation. We thus estimate the unknown null distribution function $F_{V^{(t)}_{\{1,\ldots,k\}}}$ as a Monte Carlo average over T^* replicates of t^*. This yields

$$\widehat{F}_{V^{(t)}_{\{1,\ldots,k\}}}(v) = \frac{1}{T^*}\sum_{t^*=1}^{T^*} I_{]-\infty,v]}(V^{(t^*)}_{\{1,\ldots,k\}}), \qquad [14]$$

for $v \in \mathbb{R}^+$, and

$$\hat{\zeta}_\gamma = \inf\left\{v : \widehat{F}_{V^{(t)}_{\{1,\ldots,k\}}}(v) \geq \gamma\right\}$$

for the corresponding estimate of the γ quantile. Therefore, we reject hypothesis **6** at nominal test size α, and we consider trader t a potential fraudster, if

$$v^{(t)}_{\{1,\ldots,k\}} > \hat{\zeta}_{1-\alpha}, \qquad [15]$$

where $v^{(t)}_{\{1,\ldots,k\}}$ is the observed value of $V^{(t)}_{\{1,\ldots,k\}}$.

Motivated by large-scale applications, Efron (30) describes a related methodology for empirically estimating a null distribution when the standard theoretical model (such as the NBL in the case of digit counts) does not hold. This approach uses the available data to estimate an appropriate version of the distribution of the test statistic under the null hypothesis. However, it is apparent that empirical null estimation is not directly feasible when recast in the framework of models **5** and **7**. One reason is that the method generally requires a known parametric form for the null distribution, whose parameters are then estimated from the available realizations of the test statistic. Even more fundamentally, in our applied context there is no guarantee that the proportion of genuine transactions is large for each trader, that is, that τ_t is small for each t in models **5** and **7**, thus violating a key assumption for empirical null estimation (ref. 30, p. 98).

On the other hand, the proportion of transactions that involve manipulated data and their impact on $\widehat{F}_{V^{(t)}_{\{1,\ldots,k\}}}$ is arguably small when considering the Cartesian products defined in Eq. **10**. First, both $\mathcal{U}_j$ and $\mathcal{Q}_j$ are not trader specific, since they contain all of the transactions in the market for the corresponding good, and the resulting idealized transactions are further aggregated to obtain the required basket of n_t transactions on m_t products. Second, as already reviewed in the statistical background, an intrinsic robustness property of the NBL specification of our contamination model arises from decomposition **4**, since the product of independent random variables follows the NBL if only one of the factors does, regardless of the other factors (ref. 8, p. 188). We may thus expect a reduction in the contamination effect produced by a manipulated element of $\mathcal{U}_j$ (respectively, $\mathcal{Q}_j$), after multiplication by a genuine element of $\mathcal{Q}_j$ (respectively, $\mathcal{U}_j$). Third, if the NBL does not hold, the contaminant distribution $\Upsilon_k^{(t)}(d_1,\ldots,d_k)$ for a trader t may not be too far from the genuine distribution $\Psi_{t'}(d_1,\ldots,d_k)$ for some other trader $t' \neq t$, which further reduces the degree of anomaly of the corresponding realizations in the whole market. We thus see our estimate $\widehat{F}_{V^{(t)}_{\{1,\ldots,k\}}}$ as the outcome of an extended null estimation approach, where $F_{V^{(t)}_{\{1,\ldots,k\}}}$ is estimated by exploiting all of the potential samples that could have been observed given the realized transactions in the market. Since the cardinality of this sample space is very large, we finally resort to Monte Carlo simulation for approximating the extended empirical null.

Table 5 reports the estimated sizes $\hat{\alpha}$ for different values of n_t and for $m_t = 1$, when test **15** is performed at $\alpha = 0.01$ on the same sets of $t = 1,\ldots,85,500$ idealized traders already considered in Table 1, and the Monte Carlo average in **[14]** is computed on $T^* = 10,000$ independent replicates for each value of n_t. The analysis for the case $m_t = 5$ is given in *SI Appendix*,

Table 5. Estimates of test size, P, and FPR using modified procedures 15 and 16, with $T^* = 10,000$, for different values of n_t and for $m_t = 1$

No. of transactions	Test	Uniform contamination (Eq. 12)					Dirac-type contamination (Eq. 13)			
		$\tau_t = 0$	$\tau_t = 0.5$		$\tau_t = 0.8$		$\tau_t = 0.5$		$\tau_t = 0.8$	
		$\hat{\alpha}$	P	FPR	P	FPR	P	FPR	P	FPR
$n_t = 100$	$V^{(t)}_{\{1\}}$	0.071	0.414	0.716	0.928	0.600	1	0.579	1	0.572
	Test **15**	0.010	0.000	1	0.000	1	0.850	0.167	1	0.180
	Test **16**	0.011	0.350	0.329	0.864	0.179	0.990	0.161	1	0.144
$n_t = 200$	$V^{(t)}_{\{1\}}$	0.094	0.812	0.683	1	0.630	1	0.648	1	0.634
	Test **15**	0.010	0.000	1	0.000	1	0.878	0.157	1	0.187
	Test **16**	0.012	0.678	0.213	0.934	0.182	0.998	0.153	0.992	0.175
$n_t = 500$	$V^{(t)}_{\{1\}}$	0.132	1	0.719	1	0.714	1	0.708	1	0.717
	Test **15**	0.010	0.004	0.983	0.000	1	0.776	0.173	1	0.154
	Test **16**	0.010	0.894	0.189	0.938	0.149	0.996	0.171	1	0.143

The estimated test sizes for $V^{(t)}_{\{1\}}$ are also given as a reference. The nominal test size is $\alpha = 0.01$. The number of independent idealized traders in each market configuration is $T^\dagger = 85,500$ for procedure **15** and $T^\dagger = 10,000$ for procedure **16**, P and FPR. $\varsigma = 0.05$ when computing P and FPR.

section 5. In all instances, comparison with the estimated sizes of the liberal χ^2_8 test (copied from Table 1) shows that the improvement provided by our procedure is paramount. The appropriate size is also reached when n_t grows, while m_t is kept fixed. Therefore, our approach provides a valid test of **[6]** even when the asymptotic framework does not comply with the requirements of Hill's limit theorem.

We then compute P and FPR for test **15**, under the uniform contamination model **12** and the Dirac-type contamination scheme **13**, using the same sets of $t = 1, \ldots, 10,000$ idealized traders already considered in Tables 3 and 4. For simplicity, we restrict our analysis to $\varsigma = 0.05$ and $\tau_t = 0.5, 0.8$, similar qualitative conclusions being reached in the other cases. The results are again reported in Table 5 and in *SI Appendix*, section 5, for $m_t = 1$ and $m_t = 5$, respectively. We see that test **15** can have severe difficulties in discriminating between $\mathcal{T}_{\mathrm{F}}$ and $\mathcal{T}_{\mathrm{NF}}$, unless $\Psi^{(m_t,n_t)}_k(d_1, \ldots, d_k)$ and $\Upsilon^{(t)}_k(d_1, \ldots, d_k)$ are well separated or τ_t is close to one. One reason for the observed loss of power is the large number of goods that are potentially involved in the Monte Carlo estimation process. Indeed, $m_{t^*} = m_t$ for each idealized trader t^* contributing to **[14]**, but the specific goods for which the digit distribution is obtained usually vary from trader to trader. This variability inflates the quantile estimate $\hat{\zeta}_\gamma$, especially when the ratio n_t/m_t increases.

We can obtain an improved estimate of the required quantile ζ_γ by adopting a refined version of model **7**. In this specification the genuine digit distribution depends not only on m_t, but also on the specific set of goods, say $\mathcal{G}_t$, dealt with by trader t. Consequently, we now generate the behavior of T^* idealized noncheating traders t^* with the constraint that $\mathcal{G}_{t^*} = \mathcal{G}_t$. Let $\widetilde{F}_{V^{(t)}_{\{1,\ldots,k\}}}$ denote the corresponding Monte Carlo estimate of $F_{V^{(t)}_{\{1,\ldots,k\}}}$, computed as in **[14]**. Then,

$$\tilde{\zeta}_\gamma = \inf\left\{v : \widetilde{F}_{V^{(t)}_{\{1,\ldots,k\}}}(v) \geq \gamma\right\}$$

and hypothesis **6** is rejected at nominal test size α if

$$v^{(t)}_{\{1,\ldots,k\}} > \tilde{\zeta}_{1-\alpha}. \qquad \textbf{[16]}$$

The number of ways in which a basket of m_t products can be selected out of G possible goods will be huge in any real-world scenario. Computation of $\tilde{\zeta}_\gamma$ thus becomes trader specific and cannot be automated before knowing the exact composition of $\mathcal{G}_t$, differently from $\hat{\zeta}_\gamma$, which depends only on the pair (m_t, n_t). Nevertheless, estimation time is still acceptable for routine application of the methodology. For instance, in our experiment computation of $\tilde{\zeta}_\gamma$ using $T^* = 10,000$ replicates takes on average less than 0.5 s for a trader t with $n_t = 200$ and $m_t = 5$.

The performance of the refined test procedure **16** is displayed in Table 5 (for $m_t = 1$) and in *SI Appendix*, section 5 (for $m_t = 5$). All of the estimated sizes are very close to the nominal target $\alpha = 0.01$ and similar to those obtained through **[15]**. Power values are comparable for the three reported tests when the genuine and the contaminant digit distributions are well separated. However, our proposals are still preferred since their FPR is considerably lower than for $V^{(t)}_{\{1\}}$. It is in the case of intermediate contamination, as under the uniform model, that the refined estimator $\tilde{\zeta}_\gamma$ shows much higher efficiency than $\hat{\zeta}_\gamma$. In this instance rule **16** ensures that the reduction in power with respect to the χ^2_8 test is minor, while keeping considerably lower values of FPR. We thus conclude that, having the appropriate size and power properties comparable to those of the liberal standard procedure, our modified tests **15** and **16** are recommended whenever the attained levels of FPR can be tolerated in practice.

Case Studies

To illustrate the use of the proposed procedure and its ability to detect relevant value manipulations, we first discuss the case of a trader extracted from an archive of fraudulent declarations provided by the Italian customs after appropriate data anonymization. The same archive was also used in ref. 6. The trader under scrutiny has $n_t = 648$ import transactions on $m_t = 38$ products from January 2014 to June 2015. The quantities and values appearing in the declarations of the three most traded products (not labeled for confidentiality reasons) are represented as (red) solid circles in the scatter plots of Fig. 1. The information displayed in such scatter plots is the input for some commonly adopted (robust) regression techniques aiming at the automatic detection of value frauds in customs data; see, e.g., ref. 31 and *SI Appendix*, section 7 for further details. However, the plots for this trader do not provide clear evidence of substantial undervaluation or of other major anomalies, although two of the declarations displayed in Fig. 1, *Center* were found to be fraudulent after substantial investigation. Our testing procedure instead produces a strong signal of contamination of the digit distribution. In fact, restricting for simplicity to the first digit, we obtain $v^{(t)}_{\{1\}} = 62.6$ and $\hat{\zeta}_{0.99} = 27.3$, based on $T^* = 10,000$

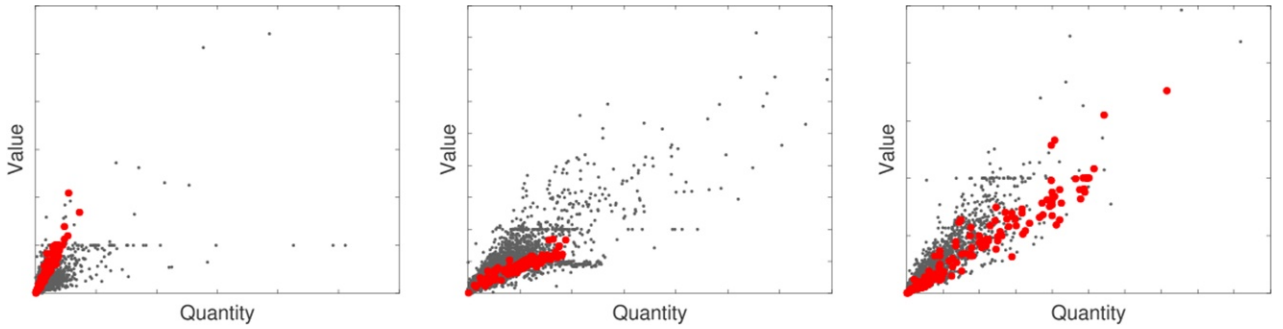

Fig. 1. Quantity-value scatter plots for the three most traded products by an Italian operator convicted for two false declarations. The transactions made by this trader are represented as (red) solid circles.

simulated traders with the same values of m_t and n_t. By applying rule **15**, we can thus conclude that hypothesis **6** can be safely rejected when the focus is shifted from individual transactions, as in Fig. 1, to the whole trader activity, as in our test.

The strength of evidence against the null may suggest the existence in the administrative records of this trader of a larger number of manipulated declarations than the two already detected. It also suggests that our method could be helpful in providing authorities with evidence of potential fraud among traders not previously classified as fraudsters or even not considered as suspicious. In view of contamination models **5** and **7**, and of our simulation results, we expect this information gain to be higher in the case of serial misconduct. Additional investigations for this trader are given in *SI Appendix*, section 6. Although all methods point to the same conclusion, we remark that simple graphical tools for conformance checking—such as histograms—require substantial human interpretation and thus cannot be routinely applied on thousands of traders.

We now move to (anonymized) data provided by the customs office of another EU member state, not disclosed for its specific confidentiality policy, that we label as MS2. The data were collected in the context of a specific operation on undervaluation, focusing on a limited set of products traded by fraudulent operators that have systematically falsified the import values. The traders classified as nonfraudulent were audited by the customs officers of MS2 and no indications of possible manipulation of import values were found. Although the absence of fraud can never be anticipated with certainty, we can thus place good confidence on these statements of genuine behavior. In *SI Appendix*, section 6 and Table S7 we provide empirical investigations of the first-digit distribution of the 15 traders in this small benchmark study for which $n_t \geq 50$, as in our simulation experiments. We apply test **16** instead of test **15**, since the available database is limited to a basket of fraud-sensitive products, and we keep $\alpha = 0.01$ and $T^* = 10,000$ for each observed pair (m_t, n_t). We give the estimated P value of each test, computed as $1 - \widetilde{F}_{V_{\{1\}}^{(t)}}(v_{\{1\}}^{(t)})$, and—as a reference—the asymptotic P value from the χ^2_8 distribution that assumes validity of the NBL. It can be seen that our approach gives very good results, both when applied to fraudsters—it clearly rejects the hypothesis of no contamination for five traders—and in the case of genuine behavior—none of the supposedly honest traders is flagged by our test at $\alpha = 0.01$. Therefore, this study supports the claim that our methodology can be an effective aid to the preparation of the audit plans of customs services, given its ability to point to potential serial fraudsters, in agreement with current guidelines for the customs modernization process (32). We finally note the beneficial effect of our correction for one supposedly honest trader shown in *SI Appendix*, Table S7, whose small basket of traded products may imply spurious deviation from the NBL when the classic χ^2_8 approximation is used. An extreme example of this effect is also shown in *SI Appendix*, section 6.

Discussion

We have developed a principled framework for goodness-of-fit testing of the NBL for antifraud purposes, with a focus on customs data collected in international trade. Our approach relies on a trader-specific contamination model, under which fraud detection has close connections with outlier testing. We have given simulation evidence, in the context of a real EU market, showing the features of the traders for which we can expect the genuine digit distribution to be well approximated by the NBL. Our simulation experiment is an empirical study addressing this issue in detail in the context of international trade, where the contrast of fraud has become a crucial task and substantial investigations are often demanding and time consuming. We have also provided simulation-based approximations to the distribution of test statistics when the conditions ensuring the validity of the NBL do not hold. These approximations open the door to the development of goodness-of-fit procedures with good inferential properties and wide applicability.

Our methodology is general and potentially applicable to any country, or year, for which detailed customs data are available. Being mostly automatic, it is suited to be implemented in large-scale monitoring processes in which thousands of traders are screened to find the most suspicious cases. It can also be a valuable aid to the design of efficient and effective audit plans. Although we expect our general guidelines to remain valid in other empirical studies, the specific quantitative findings may clearly vary from one country (year) to another.

A bonus of our contamination approach is that it makes clear the setting in which statistical antifraud analysis takes place. Our conformance testing procedures mainly aim at the detection of serial fraudsters, for which information accumulates in the corresponding transaction records. The generation of low-price clusters of anomalous transactions is a typical consequence of this cheating behavior, and robust clustering techniques can also be used for its detection (e.g., ref. 4). However, rejection of our goodness-of-fit null hypotheses often provides more compelling evidence of fraud, also because it may not be easy to identify the low-price clusters that actually correspond to illegal declarations. Testing conformance to the NBL, or to another suitable distribution for genuine digits, thus shifts the detection focus from individual transactions to the full set of data from each trader.

A word of caution concerns the fact that not all possible frauds can be detected by our method, even when we restrict to manipulation of transaction values. For instance, we cannot expect any statistical procedure (including our own proposal) to have high power against data fabrication methods that preserve the validity of the NBL, at least approximately, and against

occasional frauds for which statistical tests are not powerful enough. Therefore, we do not see our methodology as the ultimate antifraud tool, but as a powerful procedure to be possibly coupled with additional information. We support integration of the signals provided by our method with those obtained through alternative statistical techniques and with less technical model-free analyses—such as those developed in refs. 7 and 10—that can be applied on a restricted number of traders. Indeed, we see our approach as a suitable automatic tool for selecting the most interesting cases for additional qualitative and quantitative investigations, while ensuring control of the statistical properties of the adopted tests.

ACKNOWLEDGMENTS. We are grateful to Emmanuele Sordini for his contribution to the development of Web Ariadne, to Alessio Farcomeni for discussion on a previous draft, and to the reviewers for their helpful comments. The Joint Research Centre of the European Commission supported this work through the "Technology Transfer Office" project of the 2014–2020 Work Programme, in the framework of collaboration with EU member states customs and with the EU Anti-Fraud Office. This research line would not be feasible without factual collaboration of the customs services, enabled by the Hercule III Anti-fraud Programme of the European Union.

1. European Commission (2014) Operation SNAKE: EU and Chinese customs join forces to target undervaluation of goods at customs. Press release IP-14-1001. Available at europa.eu/rapid/. Accessed September 12, 2014.
2. Fogelman-Soulie F, Perrotta D, Piskorski J, Steinberger R, eds (2008) *Mining Massive Data Sets for Security: Advances in Data Mining, Search, Social Networks and Text Mining, and their Applications to Security* (IOS Press, Amsterdam).
3. Cerioli A (2010) Multivariate outlier detection with high-breakdown estimators. *J Am Stat Assoc* 105:147–156.
4. Cerioli A, Perrotta D (2014) Robust clustering around regression lines with high density regions. *Adv Data Anal Classif* 8:5–26.
5. Cerasa A, Cerioli A (2017) Outlier-free merging of homogeneous groups of pre-classified observations under contamination. *J Stat Comput Simul* 15:2997–3020.
6. Barabesi L, Cerasa A, Cerioli A, Perrotta D (2018) Goodness-of-fit testing for the Newcomb-Benford law with application to the detection of customs fraud. *J Bus Econ Stat* 36:346–358.
7. Nigrini MJ (2012) *Benford's Law* (Wiley, Hoboken, NJ).
8. Berger A, Hill TP (2015) *An Introduction to Benford's Law* (Princeton Univ Press, Princeton).
9. Miller SJ, ed (2015) *Benford's Law: Theory and Applications* (Princeton Univ Press, Princeton).
10. Kossovsky AE (2015) *Benford's Law: Theory, the General Law of Relative Quantities, and Forensic Fraud Detection Applications* (World Scientific, Singapore).
11. Diaconis P (1977) The distribution of leading digits and uniform distribution mod 1. *Ann Probab* 5:72–81.
12. Knuth DE (1997) *The Art of Computer Programming, Seminumerical Algorithms* (Addison-Wesley, Reading, MA), 3rd Ed., Vol 2.
13. Hill TP (1995) A statistical derivation of the significant-digit law. *Stat Sci* 10:354–363.
14. Varian HR (1972) Letters to the editor. *Am Stat* 26:62–65.
15. Pericchi L, Torres D (2011) Quick anomaly detection by the Newcomb-Benford law, with applications to electoral processes data from the USA, Puerto Rico and Venezuela. *Stat Sci* 26:502–516.
16. Fernandez-Gracia J, Lacasa L (2018) Bipartisanship breakdown, functional networks, and forensic analysis in Spanish 2015 and 2016 national elections. *Complexity* 2018: 1–23.
17. Tam Cho WK, Gaines BJ (2007) Breaking the (Benford) law. *Am Stat* 61:218–223.
18. Michalski T, Stoltz G (2013) Do countries falsify economic data strategically? Some evidence that they might. *Rev Econ Stat* 95:591–616.
19. Berger A, Hill TP (2011) Benford's law strikes back: No simple explanation in sight for mathematical gem. *Math Intel* 33:85–91.
20. Mebane WR, Jr (2011) Comment on "Benford's Law and the detection of election fraud". *Polit Anal* 19:269–272.
21. Klimek P, Yegorov Y, Hanel R, Thurner S (2012) Statistical detection of systematic election irregularities. *Proc Natl Acad Sci USA* 109:16469–16473.
22. Goodman W (2016) The promises and pitfalls of Benford's law. *Signficance* 13:38–41.
23. Nigrini M (2015) Detecting fraud and errors using Benford's law. *Benford's Law: Theory and Applications*, ed Miller SJ (Princeton Univ Press, Princeton), pp 191–211.
24. Durstchi C, Hillison W, Pacini C (2004) The effective use of Benford's law to assist in detecting fraud in accounting data. *J Forensic Account* 5:17–34.
25. Schatte P (1984) On the asymptotic uniform distribution of sums reduced mod 1. *Math Nachr* 115:275–281.
26. European Commission (2000) Commission Regulation (EC) No 1917/2000 of 7 September 2000 laying down certain provisions for the implementation of Council Regulation (EC) No 1172/95 as regards statistics on external trade. *EUR-Lex, Official Journal of the European Union* L 229:14–26.
27. Samaniego RM, Sun JY (2015) Technology and contractions: Evidence from manufacturing. *Eur Econ Rev* 79:172–195.
28. Fan H, Li YA, Yeaple SR (2015) Trade liberalization, quality, and export prices. *Rev Econ Stat* 97:1033–1051.
29. Cerioli A (2002) Testing mutual independence between two discrete-valued spatial processes: A correction to Pearson chi-squared. *Biometrics* 58:888–897.
30. Efron B (2010) *Large-Scale Inference* (Cambridge Univ Press, Cambridge, UK).
31. Perrotta D, Torti F (2010) Detecting price outliers in European trade data with the forward search. *Data Analysis and Classification* , eds Palumbo F, Lauro CN, Greenacre MJ (Springer, Berlin), pp 415–423.
32. World Customs Organization (2017) Message from the WCO secretary general. Available at http://www.wcoomd.org/en/media/newsroom/2017/january/message-of-the-wco-secretary-general-on-international-customs-day-2017.aspx. Accessed January 26, 2017.