The Illusion of Change: Correcting for Biases in Change Inference for Sparse, Societal-Scale Data

Gabriel Cadamuro*
University of Washington
gabca@cs.washington.edu

Ramya Korlakai Vinayak
University of Washington
ramya@cs.washington.edu

Joshua Blumenstock
University of California Berkeley
jblumenstock@berkeley.edu

Sham Kakade
University of Washington
sham@cs.washington.edu

Jacob N. Shapiro
Princeton University
jns@princeton.edu

ABSTRACT
Societal-scale data is playing an increasingly prominent role in social science research; examples from research on geopolitical events include questions on how emergency events impact the diffusion of information or how new policies change patterns of social interaction. Such research often draws critical inferences from observing how an exogenous event changes meaningful metrics like network degree or network entropy. However, as we show in this work, standard estimation methodologies make systematically incorrect inferences when the event also changes the sparsity of the data.

To address this issue, we provide a general framework for inferring changes in social metrics when dealing with non-stationary sparsity. We propose a plug-in correction that can be applied to any estimator, including several recently proposed procedures. Using both simulated and real data, we demonstrate that the correction significantly improves the accuracy of the estimated change under a variety of plausible data generating processes. In particular, using a large dataset of calls from Afghanistan, we show that whereas traditional methods substantially overestimate the impact of a violent event on social diversity, the plug-in correction reveals the true response to be much more modest.

KEYWORDS
Computational Social Science, Change Detection, Entropy Estimation, Call Detail Records

1 INTRODUCTION
Over the past decade, the increasing availability of societal-scale data has led to new approaches to social science research[5, 11, 24, 38]. In this literature, one common strain of analysis studies the human response to important geo-political events, using digital trace data as a lens into that response. For instance, [36] shows how to rapidly detect an earthquake from Twitter behaviour, [3] uses mobile phone data to study collective response to several different types of emergencies, and [39] studies rumors on social media following an oil spill, to cite just a few examples.

A common methodological challenge in such research is the issue of sampling sparsity: where the likelihood of observing any given edge in the social graph during a given period may be low and lead to inaccurate estimates of an individual-level properties. This problem is well-known and there is a rich body of work[19, 32, 35, 41, 43, 46] in both theory and application considering how to better estimate in the presence of sparsity. However, additional and previously unconsidered issues arise when this sparsity may vary over time: we call this property dynamic sampling sparsity.

Figure 1: Illustrating variations in sparsity through analysis of call records during a bomb attack in a major city. Graph (a) shows how the hourly call volume of one of the impacted cell towers experiences a very noticeable surge during the emergency. This is also apparent when mapping the tower-level call volume (b) one hour before and (c) one hour after the bombing. The color of each tower represents how abnormally high the call volume is: with red representing call volumes over 5 standard deviations from the mean.

This paper is published under the Creative Commons Attribution 4.0 International (CC-BY 4.0) license. Authors reserve their rights to disseminate the work on their personal and corporate Web sites with the appropriate attribution.

WWW’19, May 2019, San Francisco, California USA
© 2019 IW3C2 (International World Wide Web Conference Committee), published under Creative Commons CC-BY 4.0 License.
ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313722
While dynamic sampling sparsity appears in many scenarios, analyzing the impact of emergency events provides a particularly illustrative example. Almost without fail, emergencies produce an immediate spike in transaction log activity (indeed, this spike often serves as the basis for emergency event detection and prediction[8, 18, 22, 36, 50]). However, this means that the sparsity of the social networks decreases at precisely the most confounding time: in the immediate aftermath of the event. An example of the abrupt change in sparsity conditions, derived from anonymized mobile phone data from Afghanistan, in the wake of a serious emergency can be seen in Figure 1. Understanding how important metrics of mobility and social diversity are impacted by such an emergency event, without being misled by the increased volume of communication, now becomes a serious challenge.

Our contribution: This paper shows how dynamic sampling sparsity of digital trace data can systematically bias downstream statistical inferences, and proposes a plug-in correction (namely, a fix that can be applied as a pre-processing step for any existing estimator) to address this problem in particular:

- We develop a general framework to show why existing methods will systematically produce spurious discoveries. We use this framework to derive a simple statistical correction.
- We benchmark against several state of the art estimators using both real-world and simulated data, under a range of dynamic sparsity conditions. We show that our correction reliably outperforms or matches these methods under all conditions.

The rest of the paper is organized as follows. Section 2 provides necessary technical background and discusses related work. Section 3 introduces a general framework to model the problem and proves that existing methods are biased. We construct a simple plug-in correction to existing estimators that is unbiased. This correction is then put to the test in section 4, where we test its performance on both real-life and synthetic examples and then examine how this alters the conclusions of a sociological analysis. Finally, in Section 5 we discuss the pertinence of our investigation to the broader computational social science community, noting that this problem extends to many scenarios outside of emergency event analysis, and suggest further questions of both practical and statistical relevance.

2 BACKGROUND AND RELATED WORK
2.1 Measuring social phenomena with societal-scale digital trace data

A common approach to current computational social science research involves the analysis of summary statistics that are derived from societal-scale digital trace data. Though the methods we propose apply to a great many such statistics, we begin by introducing a few common metrics which will serve as a running example in the analysis that follows.

The first set of metrics summarizes network structure, a generic class of metrics equally applicable to the Twitter re-tweet graph, the DBLP citation network, or a mobile phone network. Specifically, we consider network degree (which captures the number of unique connections of each node in the network, also called degree centrality) and network entropy (a measure of the dispersion of each individual’s network). For any graph, let the number of interactions between node $i$ and node $j$ during a given time period $t$ be $c_{ij}(t)$, and the total volume of $i$’s interactions $c_i(t) = \sum_j c_{ij}(t)$.

Degree $D_i(t)$ and network entropy $H_i(t)$ of node $i$ during period $t$ are defined as,

$$D_i(t) = |\{j \mid c_{ij}(t) > 0\}$$

$$H_i(t) = -\sum_j \frac{c_{ij}(t)}{c_i(t)} \log \frac{c_{ij}(t)}{c_i(t)}$$

A second set of metrics, most relevant in networks with geomarkers, capture the characteristic travel distance or diversity of locations visited. Common examples of metrics here include location entropy [51] (defined similarly to the network entropy, but over the distribution of locations visited rather than individuals called) for diversity and radius of gyration [16] for travel distance.

These network- and location-related metrics have been used in hundreds of papers on dozens of different datasets. For instance, entropy and degree have proven informative in inference tasks ranging from estimating regional unemployment from Twitter usage[25] to predicting wealth from cell-phone records [4, 10]. Related papers show similar results for mobility metrics [6, 15, 30]. In addition to proving useful on this range of societal-scale social networks, several forms of entropy have shown usefulness in aiding visualization of the DBLP citation network [37].

2.2 Estimators and Bias

As the sheer scale of data available increases, it is important to note the growing problem of sparsity. Metrics that require large number of samples from the distribution may be confounded when the number of samples (for instance, the volume of communication) is much smaller than the support size of the distribution (e.g., the number of individuals in the true distribution). This necessitates the use of estimator functions that approximate the true underlying metric. Since we are interested in the predictive accuracy of the estimator, we focus mainly on its bias and variance (Equation (2)), the former of which underlies the problem discussed in this paper.

Definition 2.1. Let $\hat{\theta}(Y)$ be an estimator of true parameter $\theta^*$ using the data $Y$. The bias and variance of $\hat{\theta}$ is defined as,

$$ bias(\hat{\theta}) = E(\hat{\theta}) - \theta^* , \quad var(\hat{\theta}) = E[(\hat{\theta} - E(\hat{\theta}))^2].$$

Note that the expectation $E[.]$ is over the randomness of data.

It is important to note that many key social metrics, including all of those defined in Section 2.1, have serious issues with bias when being estimated. Network degree and any entropy based metric have no unbiased estimator [29]. Obtaining an unbiased estimator for the radius of gyration, since it is related to the standard deviation of locations visited, is known to be a hard problem [47].

2.3 Inferring changes

Detecting and quantifying the impact of an exogenous geopolitical event on a social metric of interest (either over time or across locations) can provide important insight into how such events impact the behavior of larger populations. Examples in the literature include using anomalies detected in social media [20, 23] and mobile phone data [22, 50] to infer the severity and location of damage from natural disasters, or the impact of employment shocks [40]. Many of these difference detection techniques transfer smoothly across data-sets: techniques first applied to social media and communications data can be adapted to a data set as dissimilar as credit card transactions [9].
Non-parametric paired tests are used to detect if there is a systematic change in the mean of a metric of interest, say $X$, over the same population before and after an event or a treatment. For example, the Wilcoxon signed-rank test takes the paired difference of $X$ before and after an event, ranks them in the order of magnitude and then uses the rank and the sign of the difference (discarding the actual magnitude of difference to avoid the effect of heavy-tail noise) to determine whether a change occurred. However, such tests (implicitly) assume that the bias in measuring in $X$ is the same before and after the event. The proportion of times a paired test detects a change when there is no actual change (null hypothesis) is called the type I error rate ($\alpha$) and when the underlying value did indeed change this proportion is called the power ($\beta$).

**Why bias matters:** In contrast to the assumptions of the paired tests, the bias in estimating quantities like entropy depend on the sparsity of the observed distribution. Since the sparsity levels can differ widely before and after an event, the bias in the measurement of entropy will also be different. Therefore, when we take a paired difference, we are not only measuring the change, but also an additional unknown bias term that is difficult to isolate. Even when there is no change in entropy, a systematic bias due to dynamic sampling sparsity can lead to a consistently increased rate of type I errors. This is discussed more formally in Section 3. The implications of dynamic sparsity on bias and consequently on the outcome of change detection is discussed Section 3.1. We systematically study this effect for state-of-the-art entropy and support size estimators in Section 4.

### 2.4 Related work

Estimating the support size, entropy and general symmetric functions of discrete distributions when the number of observations is much smaller than the support size of the distribution is a fundamental problem that has been very well-studied [2, 13, 14, 17]. It is still an active area of research in statistics, information theory as well as computer science [1, 21, 26–28, 31, 34, 41–43, 45, 46, 48, 49]. While this research has improved state-of-the-art estimators, the well as computer science [1, 21, 26–28, 31, 34, 41–43, 45, 46, 48, 49].

To give a concrete example, the optimal number of samples needed to estimate the entropy of a discrete distribution within $\epsilon$ error is $\Theta\left(\frac{S}{\log S}\frac{1}{\epsilon^2}\right)$, where $S$ is the support size of the distribution. In practice, we do not have the luxury of soliciting more samples to meet this bound and consequently the estimation of entropy per individual in a network will incur some non-negligible bias. As we will see in Section 4.1, this can lead to systematic inference errors in a way that falls outside of this body of statistical literature.

In contrast to the situation in the statistical literature, the issue of dynamic sparsity arises when analyzing social graphs. This has been an issue in particular when looking at mobility since key metrics like radius of gyration and location entropy have issues with estimator bias. Using a more densely sampled signal that is normally

---

1A function over a discrete distribution is said to be a symmetric function if it remains invariant to relabeling of domain symbols.
2The notation $h(n) = \Theta(g(n))$ means that $h$ is bounded both above and below by $g$ asymptotically.

---

**Figure 2:** Generative model for the data for a single period of time (a) when sparsity is stationary, and (b) when sparsity is non-stationary.

not available, one work [51] showed a systematic underestimation of mobility metrics using call networks that was greater for individuals making few calls. This has also been previously seen in [33] which found that while key locations were generally well inferred, functions like location entropy or radius of gyration likewise had similarly unbalanced issues with bias. However neither of these works offered general solutions to the problem. Heuristics such as dividing a biased metric by the number of communications [7] have no guarantees in improving accuracy: whether they mitigate or aggravate the problem is entirely dependent on the underlying distributions, functions and sample sparsities. A recent work [44] has analyzed the specific bias induced on location metrics by the location-varying tower density and provided correction specially designed for their specific application setting.
We propose downsampling the observed distributions to the same number of samples before estimating the difference in $f$ of the distribution $d_t$ before and after an event. For each individual $i$, let the difference be denoted by $\delta_i := f(d_i(a)) - f(d_i(b))$, where $a$ and $b$ stand for after and before respectively. However, we do not get to see the distribution $d_i$ itself, and instead we get to observe $\tilde{d}_i$ which has $c_i$ samples from the true distribution $d_i$. Given an estimator $\hat{f}$, the intuitive way to use it to find the difference $\delta_i$ is to estimate on the before and after distributions separately and then take the difference. This gives us the estimator for the change $\hat{\delta}_i := \hat{f}(\tilde{d}_i(a)) - \hat{f}(\tilde{d}_i(b))$, where $\hat{f}(\tilde{d}_i(t))$ denotes the estimate of $f$ on the observed distribution $\tilde{d}_i(t)$. Using Equation (3), the expected difference can then be written as,

$$E(\hat{\delta}_i) = \delta_i + B(d_i(a), \hat{\lambda}_i(a)) - B(d_i(b), \hat{\lambda}_i(b)).$$

Under the null hypothesis, the underlying distributions remain the same before and after, i.e. $d_i(a) = d_i(b)$. Therefore, under the null, $\delta_i = 0$. When we test for change, we want to control $\alpha$ (the chance of declaring a change when the null is true). If sparsity was stationary, i.e $\lambda_i(b) = \lambda_i(a)$, the mean of the difference would be zero since bias would cancel out when we take paired differences (Equation (4)). However, since the observed distribution also depends on the non-stationary rate parameter $\lambda_i(t)$, the mean of paired difference is not zero under the null. For $E(\hat{\delta}_i)$ to be zero under the null hypothesis, we need the following to hold, for all $d_i(a), \lambda_i(a)$ and $d_i(b)$,

$$B(d_i(a), \lambda_i(a)) = B(d_i(b), \lambda_i(b)).$$

For functions like entropy, which do not have unbiased estimators[29], such a condition would never hold for any non-trivial distribution $d_i$ and estimator $\hat{f}_i$. This leads to a systematically increased type-I error rate under classic tests like Wilcoxon signed-rank test as illustrated in Figure 3.

### 3.2 Correction by downsampling

We propose downsampling the observed distributions to the same number of samples before estimating $f$ as a plug-in solution to avoid this problem in change detection tests.

Let $c_{\text{min}} := \min(c_i(a), c_i(b))$, and $\tilde{d}_i(a, l)$ and $\tilde{d}_i(b, l)$ be obtained by drawing $l \leq c_{\text{min}}$ samples from $\tilde{d}_i(a)$ and $\tilde{d}_i(b)$ respectively. The downsampling-corrected version of estimator $\hat{f}_i$ for difference is then defined as follows,

$$\hat{\delta}_i := E\left(\hat{f}(\tilde{d}_i(a, l))\right) - E\left(\hat{f}(\tilde{d}_i(b, l))\right),$$

where the expectation is over the randomness of drawing $\tilde{d}_i \sim d_i$. In practice this can be approximated by averaging over a few random re-samplings. Downsampling ensures that under the null hypothesis, the bias in estimating $f$ is same for before and after and hence it cancels out when we take paired difference. The situation where the null hypothesis is false is significantly harder to analyze but the performance of the proposed correction in this case is explored empirically in Section 4.2.

### 4 EMPIRICAL ANALYSIS

To verify the pertinence of this problem in real-life analysis we perform a number of empirical studies. In all of these we focus on inferring the change in two metrics: social entropy and network degree. We pick these two since they are both socially informative as well as ubiquitously available over many different types of social graphs. We are interested in how estimates in the change of these metrics are impacted by the variation in sparsity, which we quantify as the elevation rate $r$.

We compare the performance of the following four estimators:

1. **Naive-Estimator**: This simply computes the metric by treating the empirical distribution as the true distribution.

2. **Jackknifed naive**[12]: This is the naive estimator with a jackknife heuristic that averages the naive estimate over all distribution generated by removing one sample from the empirical distribution.

3. **JVHW[21]**: This estimator combines an unbiased estimator for the best polynomial approximation of the function being estimated in the non-smooth region with a bias-corrected estimate on the smooth region.

4. **APML[31]**: Approximate Profile Maximum Likelihood Estimator is a computationally efficient approximation of the profile maximum likelihood [1] which maximizes the probability of the observed profile (multiplicities of the symbols observed ignoring the label).

Note that JVHW is only applicable to entropy and not network degree. We compare these estimators to their corrected variants, where we downsample the data (as described in Section 3.2) before running these estimators. We found the results broadly similar across the corrected version of these four methods. In the interest...
4.1 Natural experiments with real-world data

In the first set of experiment we use a country-wide CDR dataset collected over 6 months in Afghanistan. This data comprises data for millions of callers and since our interest concerns changes in specific groups of individuals we restricted this to calls from a set of $N = 1000$ individuals determined to be living near a specific tower in a major city. We take the empirical distribution generated by 6 months of data (with a median of 500 calls per individual) as being sufficiently well sampled to approximate the true social distributions $\{d_i\}$'s and call rates $\{\lambda_i\}$'s. We take the empirical call rates for six months and scale them down to the equivalent rate for a week $\lambda_i = \frac{r}{7} \lambda_i$. We then assign before and after distributions to be identically $d_i(a) = d_i(b) = d_i$ and $\lambda_i(a) = \lambda_i$ but we multiply the second calling rate by the elevation rate: $\lambda_i(b) = r \lambda_i$. We repeat 100 trials where we sample using these distributions and $\lambda$'s as in Figure 2(b) and compare the estimated difference between the metric average of sets $a$ and $b$. We run the Wilcoxon signed-rank test and check if it detects a change. Since the distributions are the same, ideally we would like to estimate that there is no difference. Figure 3 shows the results for social entropy and network degree: though the same trend is present in both. We clearly see that all the methods that do not correct for varying sparsity, including cutting-edge estimation techniques like JVIHW and APML, reveal substantial issues with bias at even modest elevation rates which get progressively worse as the rate increases. In contrast, our corrected method consistently returns the correct result no matter the level of imbalance in sparsity.

4.2 Synthetic tests

While experiments on real data are essential to proving the practical concerns around the sampling problem they only provide a fixed set of conditions to experiment with. For this reason we created a synthetic test suite that would allow us to compare our methods against baselines on a variety of distributions and at a significantly more granular level. This allows us to directly set $d_i(a)$ and $d_i(b)$ to both explore different distributions and also be significantly different. As such we can compute the bias of estimators when $E[\delta_i] \neq 0$ as well as for the null case where $E[\delta_i] = 0$.

The experiment then proceeds similarly to section 4.2: with the exception that $\lambda_i$ and $d_i(a)$ are drawn randomly from a prior distribution. $\lambda_i$ is consistently distributed by a log-normal with mean of 50 while we perform separate experiments where the $d_i(a)$ are drawn from a distribution of either Dirichlet (with Dirichlet parameter $\alpha_D = 1.0$), geometric (with average probability of success $p = 0.9$) or uniform distributions. For the case where we wish $E[\delta_i] \neq 0$, we additionally alter the parameter of $d_i(a)$ by some fixed amount to generate $d_i(b)$.

We note that the existing methods will have substantial bias in the null case no matter how large the population $N$ is, as shown in Figure 4(a). The variance decreases as a function $N$, but not the bias. We set $N = 1000$ for the remainder of our synthetic experiments. Figure 4(b) shows that decreasing the sparsity, or equivalently increasing the observation rate $\lambda$, of course helps all methods: though as previously noted this is rarely possible in practice.

We investigate how the estimators perform in the case of both no change and some change: a subset of our results are shown in Figure 5. Our results for the null case reinforce our conclusions in Section 4.1: there is considerable variance between the different distributions and uncorrected metrics but our correction consistently return an accurate estimate (Figure 5(a(−d))). This illustrates the difficulty of the problem when not accounting for variable sparsity: a non-corrected method that seems to work on one distribution may entirely fail on another. We also record how often the Wilcoxon signed-rank test records a true positive as a function of the actual average difference. We see that the elevation rate has induced an asymmetric change in non-corrected methods and hence worse discovery rates when the true change in entropy is negative. On the other hand, the corrected method is reliable through-out (Figure 5(e−h)). Even in a situation where a given uncorrected method perform well (notably, the APML method is fairly robust in the uniform scenario for both network degree and entropy), the corrected method has comparable or better sensitivity while outperforming it in all other situations. This provides strong evidence that the plug-in correction is an improvement also in the case where there is a difference.

4.3 Analysis of sociological events

In this section, we highlight the relevance of this problem to computation social science by demonstrating how it can alter the conclusion of a real analysis. Recalling the call dataset described in Section 4.1, we cross-referenced calls made in that set with the time and location of a serious bomb attack and generated a set of 220 individuals who appear to live in the vicinity of this attack. Our goal is now to analyze how the average network entropy changes
in the immediate aftermath of this emergency. For each 24-hour period in our date range we take the difference with respect to the same period one week before. For example the 24-hour period starting on August 22nd 10am is paired with the 24-hour period starting on August 15th 10am, the one starting at August 22nd 11am is paired with that starting on August 15th 11am etc. We then compare how different methods infer changes based on these differences: our results are shown in Figure 6. While both the basic methods and our correction to JVHW method detect an increase during the emergency period, the uncorrected methods detect anywhere from twice to three times as much of a change. Moreover, the corrected method finds only one 24-hour period to be statistically significantly different: while the other methods declare almost the entire period to show a significant increase in network-entropy.

5 CONCLUSION
This paper explains and formalizes the concept of dynamic sampling sparsity, and highlights why it is such an important problem for estimation and change detection. Our statistical framework shows that failing to account for varying sparsity in the data frequently leads to systematic errors in the downstream statistical analysis. We demonstrate the severity of this issue through experiments on both real social graph datasets and comprehensive synthetic tests.

While we motivated this problem by considering the real-world problem of understanding the impact of emergency events, we note that this problem of varying sparsity is significantly broader. Indeed the issue would likely arise when comparing average values of social metrics (whose bias gets influenced by sampling sparsity) between two different populations with different sampling sparsity rates. Examples in the literature include comparing the structure of social networks in urban locations with that of provincial villages [10], or wealthy provinces to a poorer ones [11, 25]. Our empirical results show that it is very hard to determine ahead of time how much a specific scenario will be affected: the impact is a complex function of the different sparsity rates, the underlying distributions and the estimators themselves. The correction we develop can help avoid such errors in arbitrary environments.

A broader implication of the results in this paper is that great care is needed when performing empirical analysis on societal-scale datasets with non-stationary sampling sparsity. Many common distributional tests fail when two distributions are generated from different sparsity regimes. Rather than applying one-off fixes to each such biased metric, more research is needed into optimal statistical detection, estimation and inference tools for large-scale heterogeneous and sparse datasets.
REFERENCES


