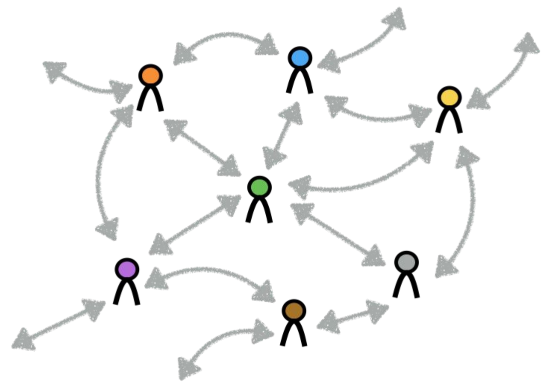
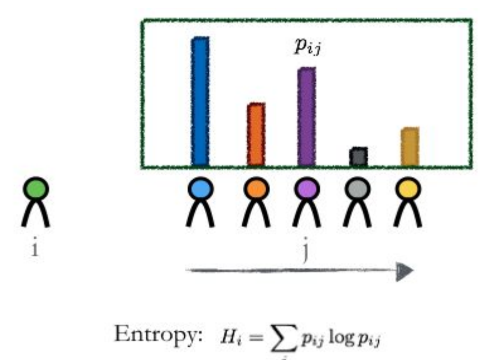
The Illusion of Change: Correcting for Biases in Change Inference for Sparse, Societal-Scale Data

Gabriel Cadamuro[*] <gabca@cs.washington.edu>, Ramya Korlakai Vinayak[*], Joshua Blumenstock[x], Sham Kakade[*], Jacob N. Shapiro[+]
[*]University Of Washington [x]University of California Berkeley [+]Princeton University

Introduction

- Computational Social Science is increasingly performed on large social networks: such as Facebook, Twitter or country-wide Call Detail Records (CDRs).
- CDRs provide both spatiotemporal metadata, as well as excellent coverage in developing countries: large body of work relating socio-economic behaviour to calling patterns.
  - Wealth and network diversity (Eagle at al 2010)
  - Phone usage as a function of social indicators (Blumenstock et al 2010)
  - Unemployment on several metrics (Toole et al 2015)
- Project that motivated this work: seek to quantify impact of violence on key social metrics using two years of CDR data on an asian country
- Which metrics matter?
  - Network degree
  - Network entropy
  - Mobility





The problem of sparsity

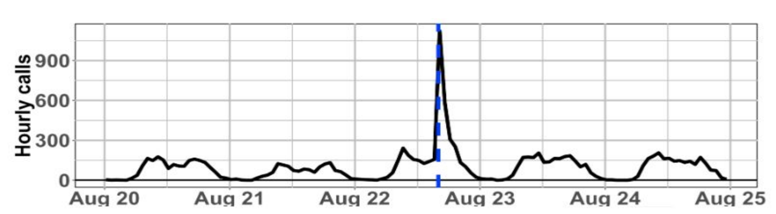
- These key metrics are functions over a discrete distribution *d*. Unfortunately, we only observe a sample of the distribution instead, this sampling sparsity introduces bias into measurements.
- Sparsity is a well-known problem with active work into mitigating bias for important functions like entropy.
- However, dynamic sampling sparsity, such as that induced by major emergency events, presents a novel source of bias that may have avoided notice.

True function value at time t: $f_i(t)$ Estimator value at time t: $\hat{f}_i(t)$

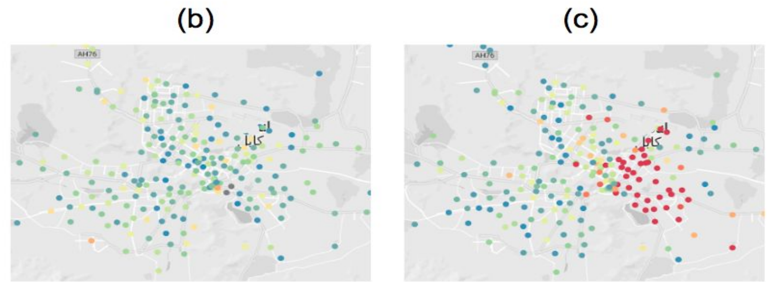
$$\text{bias}(\hat{f}_i(t)) = \mathrm{E}(\hat{f}_i(t)) - f_i(t)$$

Bias for a given estimator function depends on both distribution *d* and sample rate Lambda

$$\text{bias}(\hat{f}_i(t)) = \mathrm{E}(\hat{f}_i(t)) - f_i(t) =: B(d_i(t), \lambda_i(t))$$



(b) (c)



References

- Liam Paninski. 2003. Estimation of entropy and mutual information.Neural computation15, 6 (2003), 1191–1253.
- Jiantao Jiao, Kartik Venkat, Yanjun Han, and Tsachy Weissman. 2015. Minimax estimation of functionals of discrete distributions.IEEE Transactions on Information Theory 61, 5 (2015), 2835–2885.
- Dmitri S Pavlichin, Jiantao Jiao, and Tsachy Weissman. 2017. Approximate profile maximum likelihood.arXiv preprint arXiv:1712.07177(2017).
- Eagle, Nathan, Michael Macy, and Rob Claxton. "Network diversity and economic development." Science 328.5981 (2010): 1029-1031.
- Blumenstock, Joshua Evan, and Nathan Eagle. "Divided we call: disparities in access and use of mobile phones in Rwanda." Information Technologies & International Development 8.2
- Toole JL, Lin YR, Muehlegger E, Shoag D, Gonz´alez MC, Lazer D. Tracking employment shocks using mobile phone data. Journal of The Royal Society Interface. 2015

Model

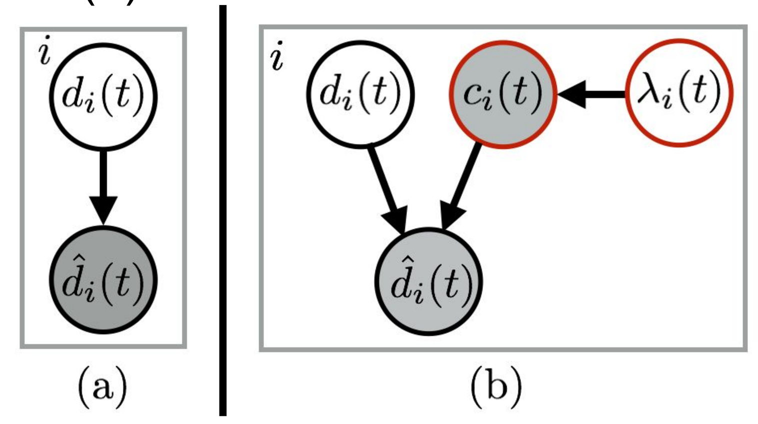
- Typically we think about the sampling process like model (a) in the figure below, the empirical distribution at time t depends only on the true distribution at time t.
- Now we must incorporate the sampling rate and interactions seen at time t as well: as shown in model (b).

Rate $\lambda_i(t)$

No. of calls seen

$$c_i(t) \sim \text{Poi}(\lambda_i(t))$$

$$\hat{d}_i(t) \overset{c_i(t)}{\sim} d_i(t)$$



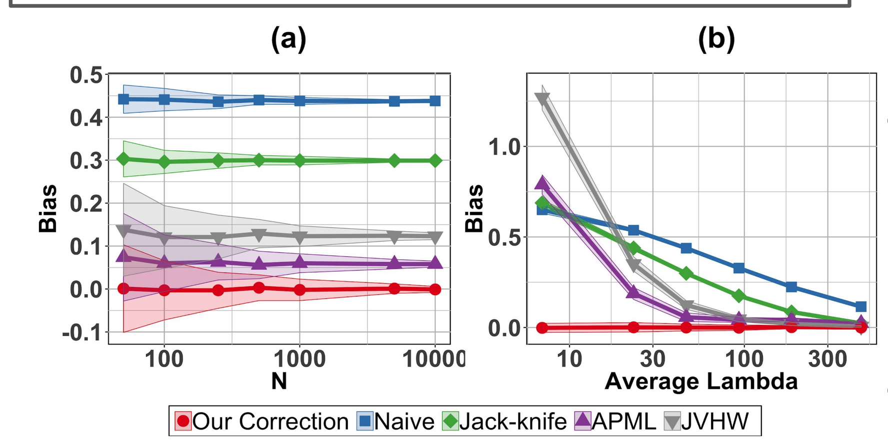
Empirical difference estimator

$$\hat{\delta}_i := \hat{f}_i(\hat{d}_i(a)) - \hat{f}_i(\hat{d}_i(b))$$

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

Necessary condition for unbiased estimate of difference

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



- Now, the bias in our estimation of the difference of a function at two points in time depends on the sampling rates at those times
- Adding more data does not mitigate this bias.
- Increasing number of samples per individual does, but in practice might not be possible/practical.
- Very real problem of False Positives when distribution is unchanged.

Our plug-in correction

- Solution works on top of any estimator (including current state of the art) for functions: simply repeatedly down-sample period with more samples to match the period with fewer samples. Then average results of estimator over all subsamples.
- Guaranteed to provide accurate results in the case of no change.
- Improves accuracy change inference in the non-null case under all conditions examined in empirical study.

Next steps

- We have addressed this problem for paired differences but there are many other types of experiments that might be affected
  - Sociological analysis comparing communities with different sparsities (e.g a village versus town)
  - Continuous time series analysis versus two period comparison
- This correction is only a partial solution, a full statistical investigation could bring several benefits
  - Estimators designed specifically with this issue in mind
  - Tight bounds as a function of elevation rate and lambda

Empirical results

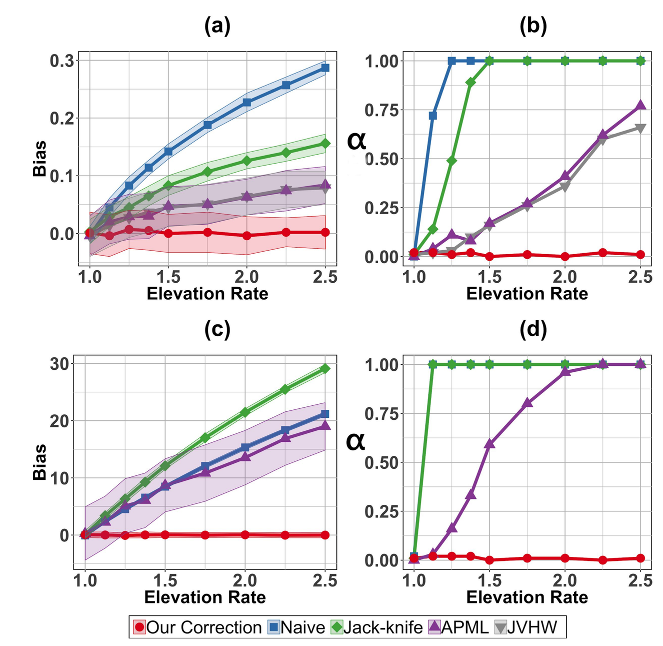
- Main goal is to quantify how change inference is impacted as a function of elevation rate *r* and verify our correction works under many conditions.

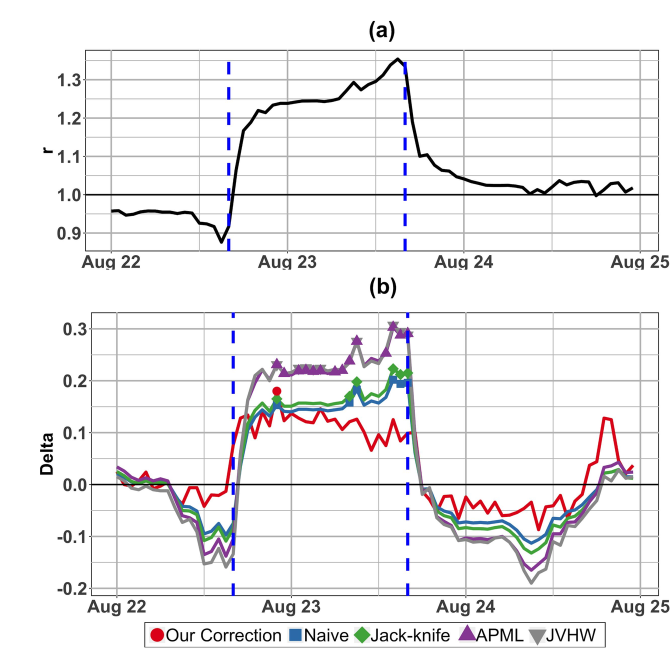
$$r := \frac{\lambda(\text{after})}{\lambda(\text{before})}$$

- Experiments will explore bias and Type I/Type II errors on Wilcox signed-rank tests
  - For both social network degree and network entropy
  - In both null (distributions do not change) and non-null cases
  - Comparing to naive/jack-knife estimators as well as state of the art estimators like JVHW and APML.

Real world data

- Drawn from 6 months of real CDR data: selects n random individuals and sub-samples full 6 months of call data at different rates to generate distribution.
- Despite being drawn from same distribution, *r* =2.0 causes even state of the art estimators to create a Type I error 50% of the time.
- Confirming theoretical results, our plug-in correction has no bias in the null scenario.
- We show that variable sampling sparsity impacts real scientific studies. The estimated change in social net entropy for a violent event was 50-100% higher when correction not applied.





Comprehensive synthetic test suite

- Synthetic tests allow us to test non-null case as well as verify results on a wide variety of base distributions
  - Base distributions: Dirichlet, uniform or geometric
  - Number of samples: lognormally distributed with mean = 50
- Results show the correction always results in a less bias, though the improvement varies as a function of distribution and function estimated.