Addressing Novel Sources of Bias for Change Detection on Large Social Networks

Gabriel Cadamuro [1] **Ramya Korlakai Vinayak** [1] **Joshua Blumenstock** [2] **Sham Kakade** [1] **Jacob N. Shapiro** [3]

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.

1. Introduction

Over the past decade, the increasing availability of societal-scale data has led to new approaches to social science research (Lazer et al., 2009; Eagle et al., 2010; Spiro, 2016; Chang et al., 2014). In this literature, one common strain of analysis studies the human response to important geopolitical events, using digital trace data as a lens into that response. For instance, (Sakaki et al., 2010) shows how to rapidly detect an earthquake from Twitter behaviour, (Bagrow et al., 2011) uses mobile phone data to study collective response to several different types of emergencies, and (Spiro et al., 2012) studies rumors on social media following an oil spill, to cite just a few examples.

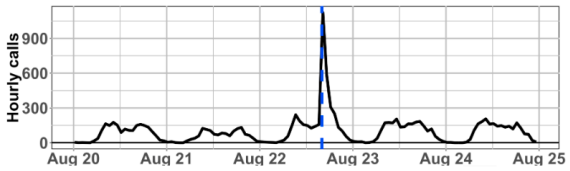

Figure 1. The extreme change in communication sparsity (measured in calls per hour) in a set of individuals in response to a violent event occurring around at the dotted blue line.

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 (Raghunathan et al., 2017; Valiant & Valiant, 2011; 2013; Saif et al., 2014; Hoteit et al., 2016) 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*.

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 (Young et al., 2014; Kapoor et al., 2010; Dobra et al., 2015; Gundogdu et al., 2016; Sakaki et al., 2010)). 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.

[1]University of Washington [2]University of California Berkeley [3]Princeton University. Correspondence to: Gabriel Cadamuro <gabca@cs.washington.edu>.

Appearing at the International Conference on Machine Learning AI for Social Good Workshop, Long Beach, United States, 2019.

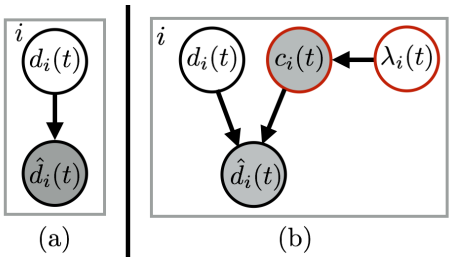

Figure 2. Generative model for the data for a single period of time when (a) sparsity is stationary, and (b) sparsity is non-stationary.

In this work we introduce and define this problem and show how it impacts a broad swathe of literature on societal-scale communication data. Using both simulated and real-life data we prove this problem materially affects the conclusions drawn from analysis. We propose a simple heuristic that is guaranteed to address the problem in some scenarios and has all-round strong empirical performance. Finally, 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

A common approach to current computational social science research involves the analysis of summary statistics that are derived from societal-scale digital trace data. Several of these statistics are often complicated functions of discrete calling distributions. Two illustrative examples are *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,

$$D_i(t) = |\{j \mid c_{ij}(t) > 0\}| \ , H_i(t) = -\sum_j \frac{c_{ij(t)}}{c_i(t)} \log \frac{c_{ij}(t)}{c_i(t)}$$

The key metrics when analyzing networks with geomarkers include *location entropy* (Zhao et al., 2016) for diversity of locations visited and *radius of gyration* (Gonzalez et al., 2008) for travel distance. The aforementioned social (Llorente et al., 2015; Cho et al., 2011; Serin & Balcisoy, 2012) and location (Pappalardo et al., 2015; Frias-Martinez & Virseda, 2012) metrics have been vital in a variety of sociological analyses on many different large social networks. Of particular interest is how changes in these metrics during emergencies (Young et al., 2014; Kapoor et al., 2010) or sudden downturns in employment(Toole et al., 2015) can be accurately tracked, for example by using paired difference tests like the Wilcoxon signed-rank test.

However it should be noted that they come with several statistical caveats. Technically we can only compute an estimator $\hat{\theta}(Y)$ for some metric of interest θ^* using the data Y available. Two key properties of an estimator in this context are $bias(\hat{\theta}) = E[\hat{\theta}] - \theta^*$, and *variance*, $var(\hat{\theta}) = E[(\hat{\theta} - E[\hat{\theta}])^2]$. Unlike simple functions like the mean of a distribution, the metrics mentioned earlier do not have an unbiased estimator: in fact there is a lively field of research looking into how to mitigate the impact of bias on function like entropy (Orlitsky et al., 2004; Valiant & Valiant, 2011; Jiao et al., 2015; Acharya et al., 2017). As such any estimate for these metrics will have some bias: the sparser the social network, the worse the bias will be.

Why bias matters: Let us explicitly consider the case where we want to track the change in a metrics like social entropy before/after an emergency event using a paired differences test. 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 a problem significantly different from the standard problem of reducing estimator bias (as in the case of entropy). While a few works empirically noted this problem in the context of location metrics (Zhao et al., 2016; Ranjan et al., 2012), they do not provide a general solution. Moreover, existing heuristic solutions such as dividing a biased metric by the number of communications (de Montjoye et al., 2016) have no guarantees in improving accuracy nor any thorough experimental analysis in this situation.

3. Theory: Dynamic sampling sparsity

In the case where sparsity is stationary, the number of samples observed before and after an event is the same on average. The generative model for the observed data is as shown in Figure 2(a), where $d_i(t)$ denotes the true distribution and $\hat{d}_i(t)$ denotes the *observed distribution* for an individual i. However, as motivated in the introduction, the sampling rate or *sparsity is not stationary* (Figure 1). In this section, we describe a general framework to capture the observation model in the setting of dynamic sparsity. Let $c \sim Poi\,(\lambda)$ denote a random variable drawn according to Poisson distribution with rate parameter λ. At time t, let $\lambda_i(t)$ be the rate of sampling for individual i and $c_i(t) \sim \text{Poi}\,(\lambda_i(t))$ denote the number of samples observed for an individual i. So, we get to observe the empirical distribution $\hat{d}_i(t)$, which is obtained by drawing $c_i(t)$ samples from the true distribution $d_i(t)$. This generative model is illustrated in Figure 2(b).

Let f be the functional (e.g: entropy) we are interested in computing on the distribution d and $\hat{f}_i(t)$ be its estimator on

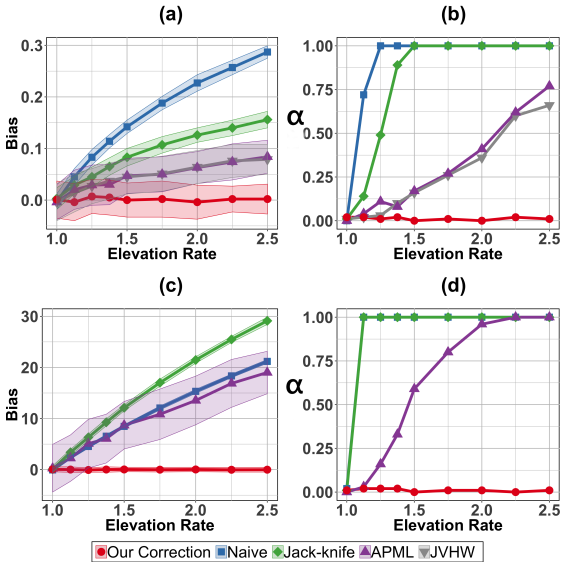

Figure 3. Comparison of how the (a) bias and (b) type-I error rate (α) for estimating difference in entropy increases with more variation in sparsity. (c) and (d) show the same for network degree. The bands in (a) and (c) show the variance in estimates.

individual i at time t. We note that this estimator is not only dependent on the true underlying distribution $d_i(t)$, but also the sampling rate $\lambda_i(t)$. Therefore, its bias can be expressed as a function B of these two:

$$\text{bias}(\hat{f}_i(t)) = \text{E}[\hat{f}_i(t)] - f_i(t) =: B(d_i(t), \lambda_i(t)). \quad (1)$$

Now consider this in the context of a paired difference test: where there are two time periods a and b and we are interested in $\delta_i := f(a) - f(b)$. In this case the corresponding estimator for δ_i is $\hat{\delta}_i := \hat{f}_i(a) - \hat{f}_i(b)$. Using this definition and equation 1 we note that

$$\text{E}(\hat{\delta}_i) = \delta_i + B(d_i(a), \lambda_i(a)) - B(d_i(b), \lambda_i(b)). \quad (2)$$

A simple corollary of equation 2 is that for $\text{E}[\hat{\delta}_i]$ to be unbiased under the null hypothesis (when $\text{E}[\delta_i] = 0$), we need the following to hold, for all $d_i(a), \lambda_i(a)$ and $\lambda_i(b)$,

$$B(d_i(a), \lambda_i(a)) = B(d_i(a), \lambda_i(b)). \quad (3)$$

For functions like entropy, which do not have unbiased estimators (Paninski, 2003), 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 both in theory and in practice (as illustrated in Figure 3).

The Downsampling correction: We provide an intuitive correction for this scenario that can be plugged into any existing estimator $\hat{f}$. Assume WLOG that $c_i(a) \geq c_i(b)$ for a given i. Then instead of using $\hat{d}_i(a)$ we generate $\widetilde{d}_i(a, c_i(b))$ by drawing $c_i(b)$ samples from $\hat{d}_i(a)$ repeatedly. We then have an estimator

$$\tilde{\delta}_i := \text{E}\left(\hat{f}(\widetilde{d}_i(a, c_i(b)))\right) - \text{E}\left(\hat{f}(\hat{d}_i(b))\right) \quad (4)$$

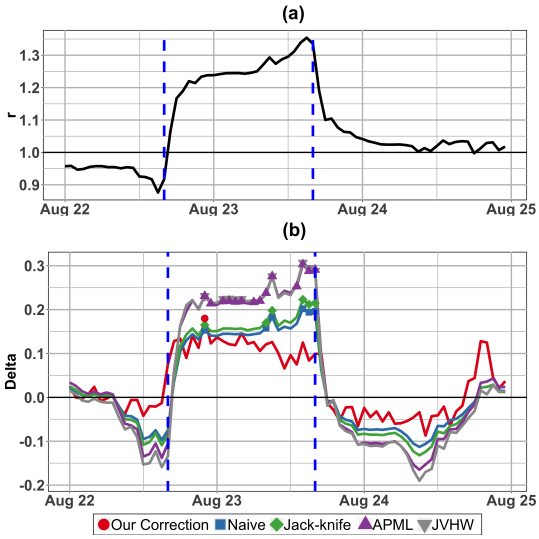

Figure 4. Analysis of (b) how different methods infer the change in network entropy in (a) the presence of varying sampling sparsity caused by a violent event. The period between the dotted blue lines indicate when the sliding window contains the bomb blast period. Marked points in (b) indicate a statistically significant difference between this 24-hour period and 24-hour period one week prior.

and vice versa for the case where $c_i(a) \leq c_i(b)$. We see that this intuitive and computationally efficient modification (since it only adds a small constant multiplier to the estimator $\hat{f}$ running time) satisfies equation 3. Consequently, it will produce unbiased results in the null case, empirical validation for this result is seen in figure 3. Providing theoretical guarantees for the non null case (when $\text{E}[\delta_i] \neq 0$) is much harder, but empirical results (figure 5) show the downsampling correction constantly improving over state of the art estimators in a variety of conditions.

4. Empirical Analysis

We perform a number of empirical studies focusing on inferring the change in network 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* $r = \frac{\lambda(\text{after})}{\lambda(\text{before})}$.

We compare our down-sampled estimator against the ubiquitous **Naive-Estimator** (simply $\hat{\delta}_i$ as in section 3) and **Jackknifed naive** (Efron & Stein, 1981) estimators as well as the state of the art **JVHW** (Jiao et al., 2015) and **APML** (Pavlichin et al., 2017) estimators. Note that JVHW is only applicable to entropy and not network degree. Our plug-in estimator can be applied to all four of the aforementioned estimators, but in the interest of clarity we only show corrected JVHW for entropy and the corrected jack-knife for network degree. The results are broadly similar in the other cases.

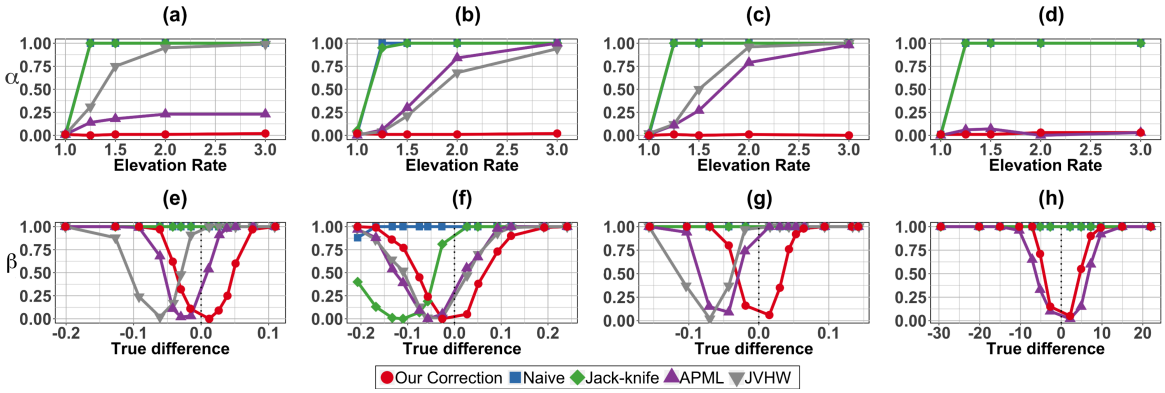

Figure 5. Panels (a)-(c): Experiments showing type-I error rates (α) for entropy change for the uniform, geometric and dirichlet scenarios respectively. Panel (d): Type-I error rate for network degree under the uniform scenario. Panels (e)-(g): Power (β) for entropy change detection at an elevation rate of 3 for the uniform, geometric and dirichlet scenarios respectively (h): Power for network degree under the uniform scenario and elevation rate of 3.

In all of these experiments we ask two questions. Firstly, what is the bias in the estimated difference for each estimator under different values of elevation rate r? Secondly, how does this translate into type-I and type-II errors? The first is simply done by computing the average predicted change and comparing it to the actual average change. The second question is studied by applying a Wilcoxon signed-rank test to the estimated differences with a desired α of 0.01.

Real CDR data: We use a country-wide CDR dataset collected over 6 months in Afghanistan comprising millions of unique callers. We focus on a set of $N = 1000$ callers living in the same area. For each trial, we subsample from the full 6 month of calls at a rate λ_a that gives the equivalent of one week's worth of calls to make period a and subsample from the same distribution at rate $\lambda_b = r\lambda_a$ to make period b. Since the distributions are the same, ideally we would like to estimate that there is no difference in either degree or entropy. As Figure 3 shows, even doubling the call frequency during period b is enough to fool state of the art estimators more than 30% of the time, while simpler estimators are totally thrown off by even $r = 1.5$. As expected, our method accurately reports no change occurs.

These results on real data are important because they bolster the assertion that accounting for sampling sparsity can material impact the conclusion of real computational social science studies. The motivating study for this work was an analysis of how metrics like social entropy change are impacted by violent events. In Figure 4 we plot how different methods and our own corrected method give substantially different results for a period of time after an event. With even a modest r of 1.2 or 1.3 the perceived change in entropy can appear to be twice or three times as high as a downsampled method perceives it to be and many more time periods appearing to show a statistically significant increase. Given the reasons to doubt non-corrected methods it is entirely possible that performing such an analysis without accounting for sampling sparsity could lead to making an incorrect inference about the effect of violence.

Synthetic data: 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. We designed a set of experiments that drew empirical call distribution as in figure 2 with a variety of base distributions $d_i(a)$ (Dirichlet with $\alpha_D = 1.0$, geometric with $p = 0.9$, and uniform) and $\lambda_i(a)$ drawn from a logNormal with mean 50. Since we can vary (and know precisely) $d_i(a)$ and $d_i(b)$, this enabled us to report empirical results for both the null and non-null ($E[\delta_i] \neq 0$) cases, which we summarize in Figure 5.This bolsters the earlier conclusions that applying the correction can only improve the accuracy of change detection since it considers both null and non-null cases as well a wider range of underlying distributions.

5. Conclusion

In this work we highlight the problem of *dynamic sampling sparsity* and show how it can seriously impact the accuracy of inferences in our setting of emergency event analysis. However, we wish to emphasize that this might be a pervasive problem in the analysis of social networks. Comparison of social metrics across two groups with living in different places (Eagle et al., 2009) or having different wealth levels (Eagle et al., 2010; Llorente et al., 2015) will also be afflicted by this problem. Studies looking at how social metrics evolve in the long term instead of just a before/after comparison will also be similarly impacted. As such, it is very important for researchers in the area to be aware of this issue and be able to estimate how much it could impact the outcome of a specific analysis. Furthermore, 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

Acharya, J., Das, H., Orlitsky, A., and Suresh, A. T. A unified maximum likelihood approach for estimating symmetric properties of discrete distributions. In *International Conference on Machine Learning*, pp. 11–21, 2017.

Bagrow, J. P., Wang, D., and Barabasi, A.-L. Collective response of human populations to large-scale emergencies. *PloS one*, 6(3):e17680, 2011.

Chang, R. M., Kauffman, R. J., and Kwon, Y. Understanding the paradigm shift to computational social science in the presence of big data. *Decision Support Systems*, 63:67–80, 2014.

Cho, E., Myers, S. A., and Leskovec, J. Friendship and mobility: user movement in location-based social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 1082–1090. ACM, 2011.

de Montjoye, Y.-A., Rocher, L., Pentland, A. S., et al. bandicoot: A python toolbox for mobile phone metadata. *J Machine Learn Res*, 17:1–5, 2016.

Dobra, A., Williams, N. E., and Eagle, N. Spatiotemporal detection of unusual human population behavior using mobile phone data. *PloS one*, 10(3):e0120449, 2015.

Eagle, N., de Montjoye, Y.-A., and Bettencourt, L. M. Community computing: Comparisons between rural and urban societies using mobile phone data. In *Computational Science and Engineering, 2009. CSE'09. International Conference on*, volume 4, pp. 144–150. IEEE, 2009.

Eagle, N., Macy, M., and Claxton, R. Network diversity and economic development. *Science*, 328(5981):1029–1031, 2010.

Efron, B. and Stein, C. The jackknife estimate of variance. *The Annals of Statistics*, pp. 586–596, 1981.

Frias-Martinez, V. and Virseda, J. On the relationship between socio-economic factors and cell phone usage. In *Proceedings of the fifth international conference on information and communication technologies and development*, pp. 76–84. ACM, 2012.

Gonzalez, M. C., Hidalgo, C. A., and Barabasi, A.-L. Understanding individual human mobility patterns. *nature*, 453(7196):779, 2008.

Gundogdu, D., Incel, O. D., Salah, A. A., and Lepri, B. Countrywide arrhythmia: emergency event detection using mobile phone data. *EPJ Data Science*, 5(1):25, 2016.

Hoteit, S., Chen, G., Viana, A., and Fiore, M. Filling the gaps: On the completion of sparse call detail records for mobility analysis. In *Proceedings of the Eleventh ACM Workshop on Challenged Networks*, pp. 45–50. ACM, 2016.

Jiao, J., Venkat, K., Han, Y., and Weissman, T. Minimax estimation of functionals of discrete distributions. *IEEE Transactions on Information Theory*, 61(5):2835–2885, 2015.

Kapoor, A., Eagle, N., and Horvitz, E. People, quakes, and communications: Inferences from call dynamics about a seismic event and its influences on a population. In *AAAI spring symposium: artificial intelligence for development*, 2010.

Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., Christakis, N., Contractor, N., Fowler, J., Gutmann, M., et al. Life in the network: the coming age of computational social science. *Science (New York, NY)*, 323(5915):721, 2009.

Llorente, A., Garcia-Herranz, M., Cebrian, M., and Moro, E. Social media fingerprints of unemployment. *PloS one*, 10(5):e0128692, 2015.

Orlitsky, A., Santhanam, N. P., Viswanathan, K., and Zhang, J. On modeling profiles instead of values. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pp. 426–435. AUAI Press, 2004.

Paninski, L. Estimation of entropy and mutual information. *Neural computation*, 15(6):1191–1253, 2003.

Pappalardo, L., Pedreschi, D., Smoreda, Z., and Giannotti, F. Using big data to study the link between human mobility and socio-economic development. In *Big Data (Big Data), 2015 IEEE International Conference on*, pp. 871–878. IEEE, 2015.

Pavlichin, D. S., Jiao, J., and Weissman, T. Approximate profile maximum likelihood. *arXiv preprint arXiv:1712.07177*, 2017.

Raghunathan, A., Valiant, G., and Zou, J. Estimating the unseen from multiple populations. *arXiv preprint arXiv:1707.03854*, 2017.

Ranjan, G., Zang, H., Zhang, Z.-L., and Bolot, J. Are call detail records biased for sampling human mobility? *ACM SIGMOBILE Mobile Computing and Communications Review*, 16(3):33–44, 2012.

Saif, H., Fernández, M., He, Y., and Alani, H. On stopwords, filtering and data sparsity for sentiment analysis of twitter. 2014.

Sakaki, T., Okazaki, M., and Matsuo, Y. Earthquake shakes twitter users: real-time event detection by social sensors. In *Proceedings of the 19th international conference on World wide web*, pp. 851–860. ACM, 2010.

Serin, E. and Balcisoy, S. Entropy based sensitivity analysis and visualization of social networks. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, pp. 1099–1104. IEEE Computer Society, 2012.

Spiro, E. S. Research opportunities at the intersection of social media and survey data. *Current Opinion in Psychology*, 9:67–71, 2016.

Spiro, E. S., Fitzhugh, S., Sutton, J., Pierski, N., Greczek, M., and Butts, C. T. Rumoring during extreme events: A case study of deepwater horizon 2010. In *Proceedings of the 4th Annual ACM Web Science Conference*, pp. 275–283. ACM, 2012.

Toole, J. L., Lin, Y.-R., Muehlegger, E., Shoag, D., González, M. C., and Lazer, D. Tracking employment shocks using mobile phone data. *Journal of The Royal Society Interface*, 12(107):20150185, 2015.

Valiant, G. and Valiant, P. Estimating the unseen: an n/log (n)-sample estimator for entropy and support size, shown optimal via new clts. In *Proceedings of the forty-third annual ACM symposium on Theory of computing*, pp. 685–694. ACM, 2011.

Valiant, P. and Valiant, G. Estimating the unseen: improved estimators for entropy and other properties. In *Advances in Neural Information Processing Systems*, pp. 2157–2165, 2013.

Young, W. C., Blumenstock, J. E., Fox, E. B., and McCormick, T. H. Detecting and classifying anomalous behavior in spatiotemporal network data. In *Proceedings of the 2014 KDD workshop on learning about emergencies from social information (KDD-LESI 2014)*, pp. 29–33, 2014.

Zhao, Z., Shaw, S.-L., Xu, Y., Lu, F., Chen, J., and Yin, L. Understanding the bias of call detail records in human mobility research. *International Journal of Geographical Information Science*, 30(9):1738–1762, 2016.