# Mobility Changes in Response to COVID-19

Michael S. Warren, Samuel W. Skillman

Descartes Labs
100 North Guadalupe Street
Santa Fe, NM 87501, USA

v1.0-2-g7fd73cb

***Abstract*—In response to the COVID-19 pandemic, both voluntary changes in behavior and administrative restrictions on human interactions have occurred. These actions are intended to reduce the transmission rate of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). We use anonymized and/or de-identified mobile device locations to measure mobility, a statistic representing the distance a typical member of a given population moves in a day. Results indicate that a large reduction in mobility has taken place, both in the US and globally. In the United States, large mobility reductions have been detected associated with the onset of the COVID-19 threat and specific government directives. Mobility data at the US admin1 (state) and admin2 (county) level have been made freely available under a Creative Commons Attribution (CC BY 4.0) license via the GitHub repository** `github.com/descarteslabs/DL-COVID-19`**.**

## I. Introduction

As of March 28, 2020, the coronavirus pandemic COVID-19 is affecting 200 countries and over 600,000 individuals [1, 2]. Efforts to prevent spread of the virus include travel and work restrictions, quarantines, curfews, cancellations and postponements of events, and facility closures. These policies aim to reduce the probability of contact between infected and non-infected persons to minimize disease transmission [3]. We hypothesize that the effect of these interventions are measurable in the aggregate geospatial statistics of people's daily movement. Measuring changes in mobility is a critical input for forecasting disease spread and assessing the effectiveness of containment strategies [4].

Geolocation reports from smartphones and other mobile devices offer a mechanism to sample the movement of individuals. There are significant privacy concerns related to the availability of such data [5]. Standard practice is to make such traces anonymous, where the true identities of nodes are replaced by random identifiers. However, the privacy concern remains. Nodes are open to observations in public spaces, and they may voluntarily or inadvertently disclose partial knowledge of their whereabouts [6]. Some studies have shown that sharing anonymized location data can lead to privacy risks and that, at a minimum, the data needs to be coarse in either the time domain or the space domain [7]. We do not and will not use mobile device location data to identify individuals. All analysis we perform is statistically aggregated, removing the ability to characterize the behavior of any single device.

## II. Method

Cloud computing with associated high-bandwidth storage capacity, combined with recent advances in machine learning, is enabling understanding of the world at a scale and level of granularity never before possible [8]. Mobile devices exemplified by the smartphone are an obvious result of this technological advance. In this work, we analyze a commercially available mobile device location dataset using cloud computing resources. We heavily leverage our festivus virtual file system layer [9] for this analysis, reading over 50 TB of input data and utilizing about 50,000 CPU hours of computation.

Raw location data from commercial data providers is generally provided in a compressed, comma-delimited text format, with one position report per line. Reports contain, at a minimum, an anonymized node id, epoch time (seconds elapsed since 00:00:00 UTC on January 1, 1970), latitude, longitude and an estimate of position accuracy. Data is obtained in multiple files to facilitate parallel processing, and a typical volume of compressed data is 100 GB per day. Data is delivered once per day.

Our analysis algorithm is implemented with the Python programming language [10, 11] in a parallel task-based execution environment. The first step in analysis is to eliminate reports with estimated position accuracy that exceeds some threshold. We use a threshold of 50 meters. Median reported accuracy across all position reports is in the range of 15-20 meters.

In order to make daily statistics more interpretable, the reported timestamp is converted to local time. If a time conversion is not performed, then the statistics in each area would be analyzed in UTC, and would not correspond to a "day" in the local timezone. For instance, a day in UTC from 00:00 to 23:59 would correctly correspond to a local day (midnight to midnight) in the UK, but would be noon to noon in New Zealand. This would lead to analysis artifacts (e.g. a day that combines Sunday afternoon with Monday morning in New Zealand cannot be consistently compared to midnight Sunday through midnight Monday in the UK). Our initial conversion to local time uses an approximate local solar time with a time zone offset of,

$$tzhours = \left\lfloor \frac{longitude}{15} \right\rceil$$

when processing individual position reports, but later determines the precise local timezone once for each device.

To facilitate the computation of statistics for each node, the reports for each node are collated. Device reports are typically not delivered in a form that can be easily processed per node, so we perform the equivalent of a parallel bucket sort [12]. Our previous work developing algorithms for computational cosmology demonstrated the utility of abstracting the most challenging aspect of a parallel algorithm into a sort [13].

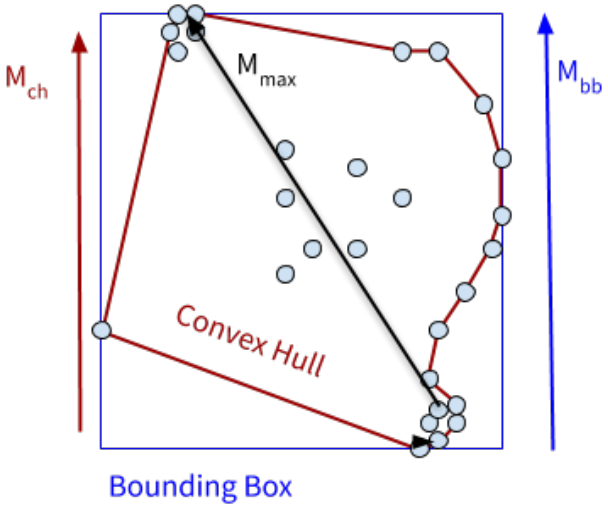


Fig. 1. An illustration of the three mobility measures we calculate from a set of positions. $M_{max}$ is the maximum distance to a point from the initial point of the day. $M_{bb}$ is the linear measure of the bounding box around the points. $M_{ch}$ is the linear measure of the convex hull of points.

We eliminate nodes that have fewer than 10 reports in a day. We also eliminate nodes that report over a short part of the day (less than 8 hours) to avoid biasing the mobility stastistic downward with transient reports. We calculate a number of statistics intended to capture the mobility of a population (Figure 1). The max-distance mobility $M_{max}$ is the maximum Haversine (great circle) distance (in units of $km$) from the initial location report of the day. Outliers (which can occur with bad position reports caused by poor GPS fixes or other errors) are trimmed by eliminating the top 10% of the distribution. The bounding box mobility is computed by determining the area of the bounding box of the reports $A_{bb}$, and then converting to an equivalent linear distance in $km$, $M_{bb} = 111\sqrt{A_{bb}}\cos(latitude)$. The convex hull mobility is computed by determining the area of the convex hull $A_{ch}$ of the reported positions, and then converting it to an equivalent linear distance in $km$ via $M_{ch} = 111\sqrt{A_{ch}}\cos(latitude)$.

For a canonical position of each processed node (such as the first or last report of the day in local time, or a derived location such as the centroid of node locations) we reverse geocode the location to a country and administrative region. We use ISO 3166-1 alpha-2 codes for the `country_code` and GeoNames feature names and codes for subcategories. `Admin1` is the first-order administrative division, being a primary administrative division of a country, such as a state in the United States. `Admin2` is a second-order administrative division, such as a county or borough in the United States. `Name` is a populated place feature, representing a city, town, village, or other agglomeration of buildings where people live and work.

For each area of interest found by the reverse geocoding procedure above, we compute a variety of statistics across the mobility measures $M_{max}, M_{bb}, M_{ch}$. These statistics include the mean, median and quartiles. For the results reported here, we focus on the median of the max-distance mobility $M_{max}$, which we designate $m50$. We further define a normalized mobility index,

$$m50\_index = 100\frac{m50}{m50_{norm}},$$

where $m50_{norm}$ is a "normal" value of $m50$ in a region, defined as the median $m50$ in that region during a designated earlier time period. We use the median weekday value of $m50$ between the dates of 2020-02-17 and 2020-03-07 as $m50_{norm}$ to investigate COVID-19 related changes in the US. Note that $m50$ has dimensions of kilometers, while $m50\_index$ is dimensionless. $m50\_index$ is easily transformed to a percentage change from baseline behavior via

$$pct\ change = m50\_index - 100.$$

Output of our analysis is stored in two formats. The first is newline-delimited JavaScript Object Notation (NDJSON) (also known as JSON lines format) which is a convenient format for storing structured data that may be processed one record at a time [14]. It works well with unix-style text processing tools and shell pipelines, and can be read into a python dictionary data structure with a simple motif

```
for line in file:
    d = json.loads(line)
```

Using a format such as the GeoJSON FeatureCollection object is problematic for very large datasets, because the file must be parsed as a whole, and records cannot be processed one-at-a-time in a streaming fashion. We also store the data in a comma-seperated values (CSV) format which may be more familiar to users of traditional geographic information systems or mapping tools.

One significant caveat to the accuracy of the statistics we present here is related to systematic sampling errors. We assume that the sampling of devices is "fair" and represents the behavior of the population as a whole. Since data is available for only a small fraction of the total number of devices (a few percent at most), it is possible that those position reports that are selected for inclusion in the raw location data we use as input have a correlation with behavior that is not representative of the average. For instance, if the data vendor includes position reports from a transportation "app", we expect those reports to show considerably more distance traveled than an app that would be used while at home. As long as the data from both sets of apps is consistently included, the systematic

sampling error would be expected to be small. However, if the data vendor loses access to one type of app, that would skew the mobility statistics, and not represent a change in actual behavior. Similar types of error would be expected if there was a correlation between changes in mobility, and the probability of appearing in the location dataset (e.g. if people who travel long distances are more likely to disable location reporting). We mitigate the effects of these sampling errors to the extent possible by analysis of multiple datasets, and we seek to validate the statistics through independent observations.

## III. Results

On January 23, 2020 the city of Wuhan was closed by canceling planes and trains leaving the city, and suspending buses, subways and ferries within it. By this date at least 17 people had died and more than 570 others had been infected, including in Taiwan, Japan, Thailand, South Korea and the United States. Ideally, we could analyze mobility around the epicenter of the outbreak in China, but we do not have access to data from that country. A recent study from researchers in Beijing was able to access aggregated statistics provided by the two largest telecommunications operators in China [15].

Between January 24 and 27, there is a dramatic drop in mobility observed in Hong Kong and Singapore 2. On Friday, Jan. 24 $m50$ in Singapore was 3.6 km. By Monday, Jan. 27 it had fallen more than 50% to 1.5 km. The change was even more pronounced in Hong Kong, dropping to 180 meters from 2.3 km. The weekend of January 25 also corresponded to the Chinese New Year holiday, so the change in mobility in that time period could have been independently influenced by the holiday.

The behavior of the these graphs for some countries between January 29 and February 27 is subject to the systematic sampling error we described previously, which shows an anomalous drop in mobility during the month of February. Using an alternate input dataset, the behavior in Singapore and Malaysia is much more consistent over that time period 3. Other notable features from that graph are the weekly cycles, with Friday having the highest mobility, particularly in the US, a lower value on Jan. 20 (the Martin Luther King Jr. Day holiday in the US) and the dip on Feb. 25 in Indonesia, which corresponds to a day of heavy rain which flooded Jakarta, the capital. Figures 2-3 also show the significant response in the US beginning March 14, and corresponding drops in the rest of the world as a new wave of mobility restrictions were put into place. Note that we do not plot values on Saturday and Sunday for clarity, since mobility on Sunday is usually much lower than weekdays.

In Figure 4 we show mobility in six selected US states beginning March 2. We detect that more rural states (with presumably longer travel distances) normally have a higher $m50$ than more urban states, specifically values of 9-11 km for Texas vs 5 km for New York for the first week in March. Reductions in mobility are detected across all six states March 13-16, with New York falling from 5.2 km on Monday, March 2 to only 31 meters on Monday, March 23. This implies

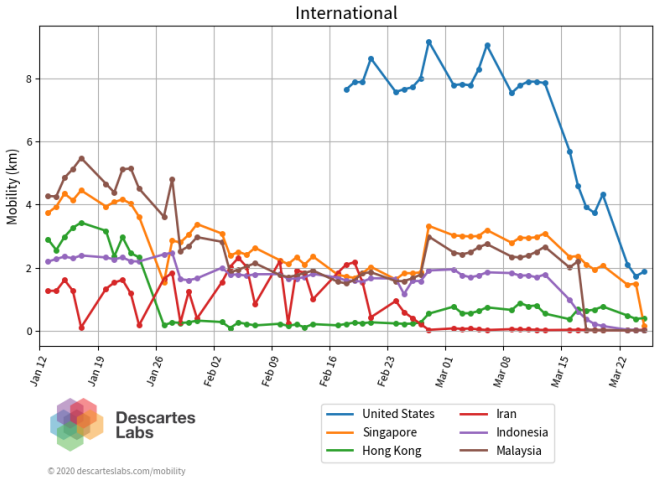


Fig. 2. Change in mobility internationally since mid-January for select countries. Between January 24 and 27, there is a dramatic drop in mobility observed in Hong Kong and Singapore corresponding to the time the spread of COVID-19 was recognized. See also the next figure for an alternate view of the same countries.

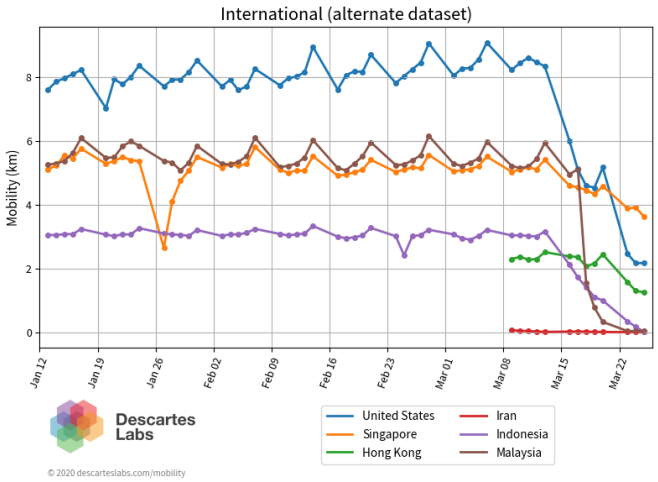


Fig. 3. Change in mobility internationally since mid-January for select countries, using an alternate dataset. The drop and recovery in Singapore on Jan. 27 is more clearly represented. Peaks on Friday are seen in the US and Malaysia, but less so in Indonesia. The drop on Feb. 25 in Indonesia corresponds to a day of heavy rain and flooding.

that half of the individuals in the entire state of New York spent most of their day within less than 100 feet of their initial position. To better understand the change from "normal" mobility and compare states with each other, Figure 5 plots the normalized $m50\_index$, demonstating that Florida and Texas have dropped to about 30% of normal, and California, Illinois, New York and Washington to less than 20% of normal.

We can also look at smaller geographic regions by plotting the change in mobility within counties. In Figure 6 we show the 10 most highly sampled counties in California. The inital reduction seen on Mar. 17 in Alameda, Contra Costa and Santa Clara counties can be attributed to the shelter-in-place order

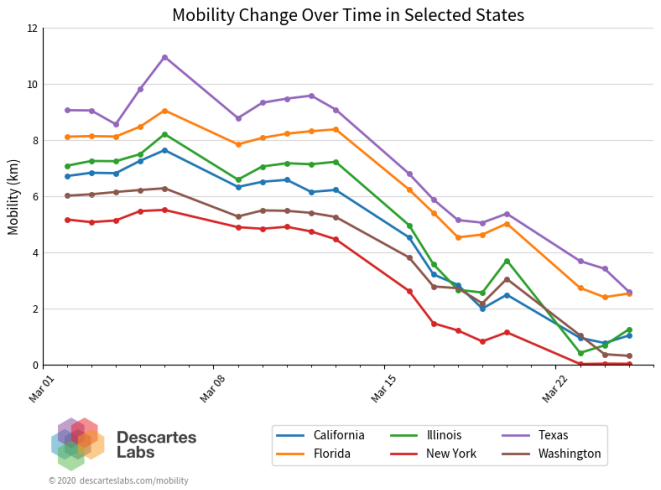


Fig. 4. Change in mobility over the first three weeks of March, 2020 in selected states in the US. Drops in mobility are detected across all six states March 13-16, with New York falling from 5.2 km on Monday, March 2 to only 31 meters on Monday, March 23. This implies that half of the individuals in the entire state of New York spent most of their day within less than 100 feet of their initial position.

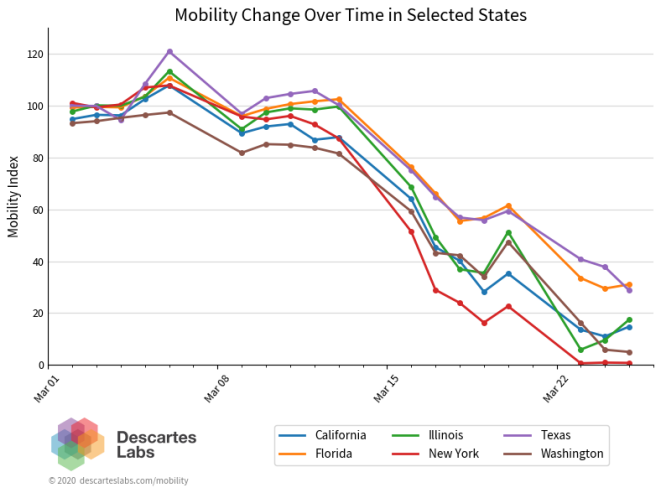


Fig. 5. Change in mobility index (normalized to 100 for the weeks of Feb. 17 through Mar. 7 2020). Florida and Texas have dropped to about 30% of normal, and California, Illinois, New York and Washington to less than 20% of normal.

for the six Bay Area counties which took effect at 12:01am. The order expanded statewide the evening of Mar. 19.

Counties in Illinois are illustrated in Figure 7. The spike in mobility on March 13 in Champaign County we believe is associated with a large number of students leaving the University of Illinois Urbana-Champaign after the on-campus classes were cancelled. A similar jump is seen in Figure 9 on March 13 in Monongalia County, associated with students leaving West Virginia University (enrollment of 30,000 students in a county with about 100,000 people). West Virgina was the last state in the US to report a coronovirus infection.

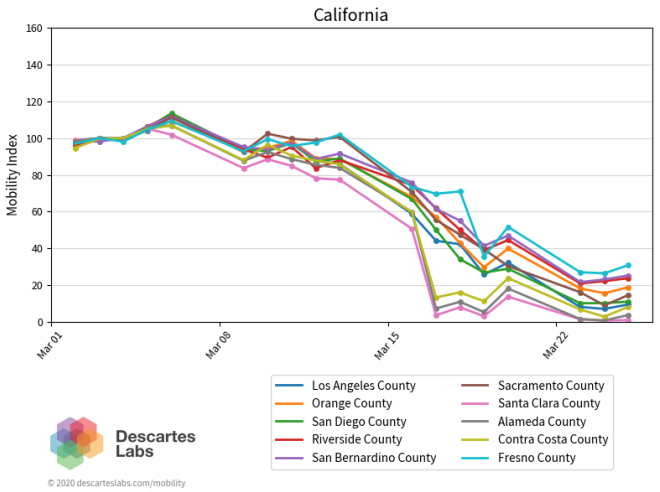


Fig. 6. Change in mobility index for the largest counties in California. The inital reduction seen on March 17 in Alameda, Contra Costa and Santa Clara counties can be attributed to the shelter-in-place order for the six Bay Area counties which took effect at 12:01am on that day. The order expanded statewide the evening of March 19.

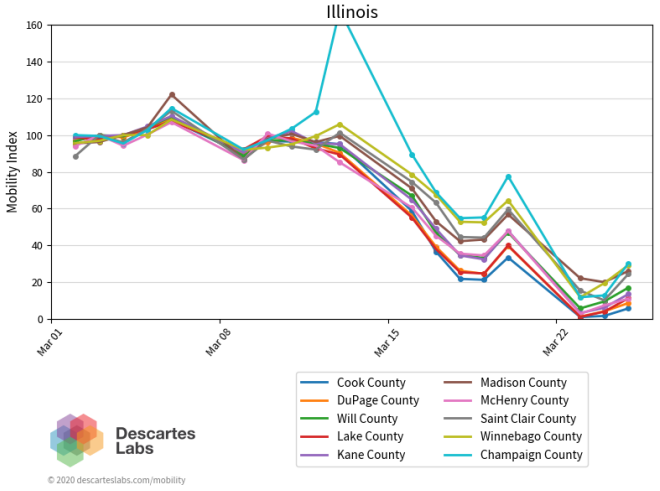


Fig. 7. Change in mobility index for the largest counties in Illinois. The spike in mobility on March 13 in Champaign County we believe is associated with a large number of students leaving the University of Illinois Urbana-Champaign after the on-campus classes were cancelled.

## IV. Conclusion

We have detected dramatic changes in mobility due to COVID-19, both within the US and globally. Mobility data at the US admin1 (state) and admin2 (county) level have been made freely available under a Creative Commons Attribution (CC BY 4.0) license via the GitHub repository `github.com/descarteslabs/DL-COVID-19`. As the situation develops, combining information such as that we derive here with pandemic growth rates in various geographies will allow more accurate models of the interventions being made, and help save lives.

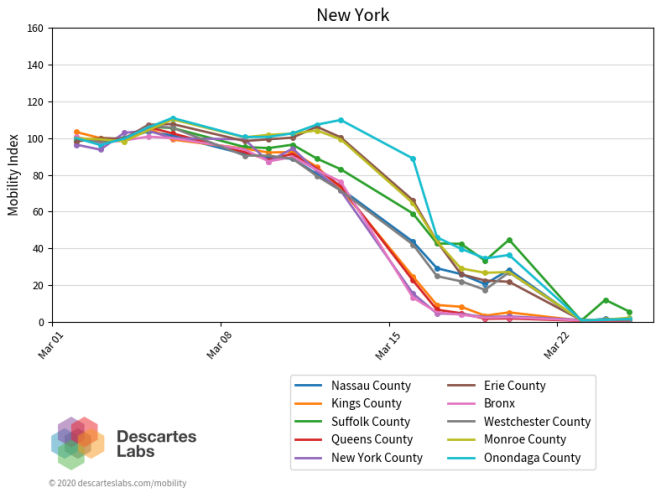


Fig. 8. Change in mobility index for the largest counties in New York.

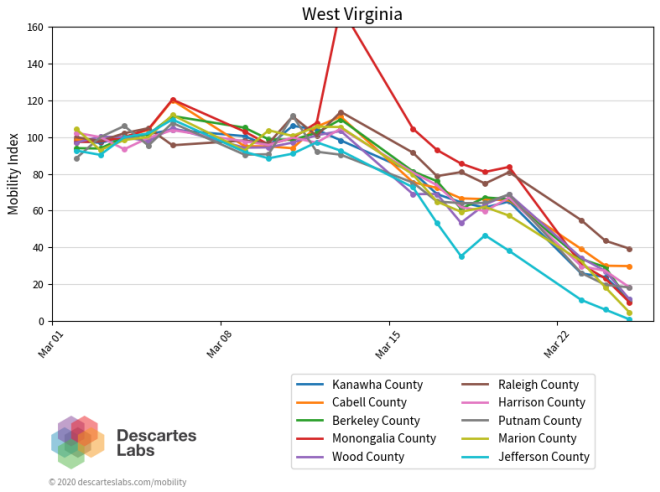


Fig. 9. Change in mobility index for the largest counties in West Virginia. The jump on March 13 in Monongalia County is associated with students leaving West Virginia University (enrollment of 30,000 student in a county with about 100,000 people).

## V. References


[1] N. Zhu et al. A Novel Coronavirus from Patients with Pneumonia in China, 2019. *New England Journal of Medicine*, 382(8):727–733, Feb. 2020. URL https://doi.org/10.1056/NEJMoa2001017. Publisher: Massachusetts Medical Society _eprint: https://doi.org/10.1056/NEJMoa2001017.

[2] World Health Organization. Novel Coronavirus (2019-nCoV) situation reports, Mar. 2020. URL https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports. Library Catalog: www.who.int.

[3] S. M. Kissler, C. Tedijanto, M. Lipsitch, and Y. Grad. Social distancing strategies for curbing the COVID-19 epidemic. *medRxiv*, page 2020.03.22.20041079, Mar. 2020. URL https://www.medrxiv.org/content/10.1101/2020.03.22.20041079v1. Publisher: Cold Spring Harbor Laboratory Press.

[4] P. Liu, P. Beeler, and R. K. Chakrabarty. COVID-19 Progression Timeline and Effectiveness of Response-to-Spread Interventions across the United States. *medRxiv*, page 2020.03.17.20037770, Mar. 2020. URL https://www.medrxiv.org/content/10.1101/2020.03.17.20037770v1. Publisher: Cold Spring Harbor Laboratory Press.

[5] C. Warzel and S. A. Thompson. Opinion | Twelve Million Americans Were Tracked Through Their Phones. *The New York Times*, Dec. 2019. URL https://www.nytimes.com/2019/12/19/opinion/tracking-phone-data.html.

[6] C. Y. Ma, D. K. Yau, N. K. Yip, and N. S. Rao. Privacy vulnerability of published anonymous mobility traces. In *Proceedings of the sixteenth annual international conference on Mobile computing and networking*, MobiCom '10, pages 185–196, Chicago, Illinois, USA, Sept. 2010. Association for Computing Machinery. ISBN 978-1-4503-0181-7. URL https://doi.org/10.1145/1859995.1860017.

[7] H. Zang and J. Bolot. Anonymization of location data does not work: a large-scale measurement study. In *Proceedings of the 17th annual international conference on Mobile computing and networking*, MobiCom '11, pages 145–156, Las Vegas, Nevada, USA, Sept. 2011. Association for Computing Machinery. ISBN 978-1-4503-0492-4. URL https://doi.org/10.1145/2030613.2030630.

[8] M. S. Warren et al. Seeing the Earth in the Cloud: Processing one petabyte of satellite imagery in one day. In *2015 IEEE Applied Imagery Pattern Recognition Workshop (AIPR)*, pages 1–12, Oct. 2015. ISSN: 2332-5615.

[9] M. S. Warren, S. W. Skillman, R. Chartrand, T. Kelton, R. Keisler, D. Raleigh, and M. Turk. Data-Intensive Supercomputing in the Cloud: Global Analytics for Satellite Imagery. *arXiv:1702.03935 [cs]*, Feb. 2017. URL http://arxiv.org/abs/1702.03935. arXiv: 1702.03935.

[10] G. Van Rossum and F. L. Drake. *Python library reference*. Centrum voor Wiskunde en Informatica, 1995.

[11] T. E. Oliphant. Python for Scientific Computing. *Computing in Science Engineering*, 9(3):10–20, May 2007. Conference Name: Computing in Science Engineering.

[12] E. Solomonik and L. V. Kalé. Highly scalable parallel sorting. In *2010 IEEE International Symposium on Parallel Distributed Processing (IPDPS)*, pages 1–12, Apr. 2010. ISSN: 1530-2075.

[13] M. S. Warren. 2HOT: An improved parallel hashed oct-tree N-Body algorithm for cosmological simulation. In *SC '13: Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis*, pages 1–12, Nov. 2013. URL https://arxiv.org/abs/1310.4502. ISSN: 2167-4337.

[14] T. Hoeger, C. Dew, F. Pauls, and J. Wilson.

ndjson/ndjson-spec, Mar. 2020. URL https://github.com/ndjson/ndjson-spec. original-date: 2013-07-05T15:50:04Z.

[15] Z. Cao et al. Incorporating Human Movement Data to Improve Epidemiological Estimates for 2019-nCoV. *medRxiv*, page 2020.02.07.20021071, Feb. 2020. URL https://www.medrxiv.org/content/10.1101/2020.02.07.20021071v1. Publisher: Cold Spring Harbor Laboratory Press.