Using network paths to find FB aggregate spending

Pavel Oleinikov

4/22/2020

Executive summary

This report explains the method used by WMP to compute the aggregate spends by regions for Facebook advertisements. The method involves solving an optimization problem - finding a shortest path on a network where the dates are nodes and reports are links. We explain why we choose a method that may have more rounding-off errors but has better readability, over the more precise method that occasionally may produce negative spend amounts.

Background

Available reports

Currently, Facebook provides several reports (https://www.facebook.com/ads/library/report/) showing how much money each advertiser has spend in a specific region:

- Daily reports. The date of the report corresponds to the day which activity is recorded. For example, today (04/22/2020), the latest available report is for `04/17/2020` and it covers the spending on that day - the 17th of April, 2020. This date is included into the name of the file that Facebook offers for download.
- Week-long reports. The date of the report is the last day of the 7-day period whose activity is covered. The `04/17/2020` report covers the period from 04/11/2020 until 04/17/2020.
- 30 day reports. Similar to the above, but covering the activity during the 30-day span. The `04-17-2020` report covers the period from 03/19/2020 to 04/17/2020.
- 90 day reports. Same logic as above. The `04/17/2020` report covers the period from 01/19/2020 to 04/17/2020.

Wesleyan Media Project (WMP) has been collecting the regional tables from daily and weekly reports since October 2019. The regional tables from the 30-day and 90-day reports were added to the collection in mid-February 2020.

Rounding-off errors in Facebook numbers

The spending numbers of small and large advertisers are reported differently by Facebook.

For advertisers whose spending exceeded $100 over the reported period, Facebook provides the exact spend, rounded off to a dollar. For example, a report may say that page named ‘X’ has spent 105 USD on advertising in California.

For advertisers whose spending is below $100, Facebook only includes a line - “< 100” - less or equal than 100 USD. The actual spend could have been anywhere between 1 USD and 100 USD - a wide margin for error.

Given that FB pages continue to spend money over time and their total numbers keep increasing, the longer the time span included into a report, the more likely a page is to exceed the 100 USD threshold, after which Facebook will report the exact amounts. A 90-day report would have substantially fewer rounded-off entries than a 1-day or 7-day reports.

Arriving to an aggregate number via multiple paths

Let's start with a hypothetical problem: I want to calculate the aggregate spend, by regions, over the time span from 04/15/2020 until 04/17/2020. The aggregations would include numbers from 3 days: 04/15, 04/16, and 04/17. I have daily reports covering activity on these dates, I take them and sum them up. I arrive at the result by summing **three** 1-day reports.

Now, let's complicate the problem. Let's say the time period is from 04/12/2020 to 04/17/2020 - six days. Because I have at my disposal both the 1-day and 7-day reports, I actually have two possible solutions:

1. Use 1-day reports: take reports for 04/12, 04/13, 04/14, 04/15, 04/16, and 04/17, and add them up. This would involve summing **six** 1-day reports.
2. Take the 7-day report posted on 04/17/2020. It covers the period from 04/17/2020 back to 04/11/2020. Then, take the daily report for 04/11/2020 and subtract its numbers from the 7-day report. Thus, I should have the aggregate numbers for the period from 04/12 to 04/17, and I obtain them using only two reports: the noisy 1-day report and the less noisy 7-day report.

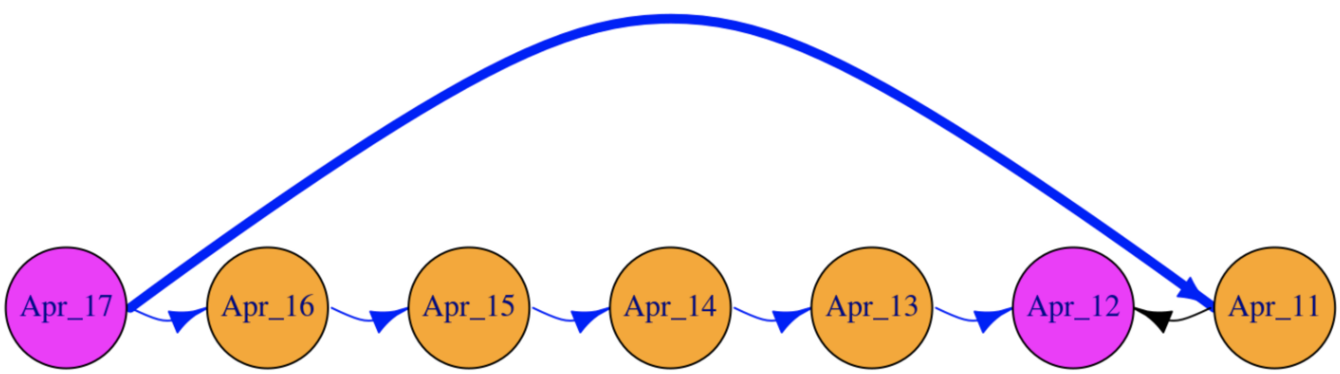

Figure 1. The edge color corresponds to the arithmetic operations: addition is blue and subtraction is black.

Assuming that the rounding-off error is plus/minus 50 USD, for some pages option 1 would give us numbers that are 300 USD off the mark. Option 2 uses only one 1-day report, so the error would be only 50 USD.

Aggregate numbers and networks paths

The above example and its two options brings up an analogy with driving directions and choosing the optimal path: with driving, there is a tradeoff between distance, speed, and potential toll fees. In the case of reports, the fee - a round-off error - is inversely related to the time span of the report.

With this idea in mind, we can formulate the problem of computing the aggregate spend as a problem of finding an optimal path on a network. In this network, in simplistic terms, the calendar dates are the nodes and reports are the edges linking them. The round-off error is the cost/penalty associated with an edge, and we want to find a path between two nodes/dates that carries the smallest penalty. In the example above, we had one path which involved six reports with a cost of 6 units (assuming that the cost is $\frac{1}{report\ time\ span}$) and the other path with a cost of $1\frac{1}{7}$.

As a further analogy with driving, we know that on some days Facebook reports are unreliable, and so these nodes are unusable. As a side note, we have a separate project that compares the daily spending reports to the differences in lifetime spending reports to identify aberrations. The most notable example of the “bad report day” was December 7, 2019, which was noticed all around the world. This CNN story (https://www.cnn.com/2019/12/11/tech/facebook-political-ads-uk-election-ge19/index.html) gives one account of the incident. When we are aware of the problematic days, we exclude their reports from the list of possible edges/links.

Below is the solution to the problem of making a path between January 1st, 2020, and April 17, 2020 - the latest currently reported date.

```
## # A tibble: 6 x 4
##   report_date report_span_starts_on  span operation
##   <chr>       <date>                <int> <chr>
## 1 2020-04-17  2020-04-17                1 plus
## 2 2020-04-16  2020-04-16                1 plus
## 3 2020-04-15  2020-03-17               30 plus
## 4 2020-03-16  2019-12-18               90 plus
## 5 2019-12-24  2019-12-18                7 minus
## 6 2019-12-31  2019-12-25                7 minus
```

In an ideal world, this would be the end of the story, however, in our case the reality of Facebook’s unreliable numbers ruins the picture. Because we have subtraction, some of the numbers come out negative. This may be fine for us, since we know the underlying process, but can be confusing to an uninformed reader.

After we generate the summary table and do the required additions and subtractions, we end up with negative numbers for a few entities. This suggests that the numbers in the 7-day reports from December were not matched in the 90-day report, since it is the one whose time span covers part of December.

Total number of rows in the aggregate spend table:

```
## [1] 588282
```

Number of rows that have negative `amt_spent`:

```
## [1] 31426
```

What is the total number of entities in the report:

```
## [1] 36235
```

How many entities have negative amounts:

```
## [1] 1535
```

What are the worst cases of negative amounts:

```
## # A tibble: 10 x 4
##    page_name            disclaimer                        region       amt_spent
##    <chr>                <chr>                             <chr>            <dbl>
##  1 Shen Yun             These ads ran without a disclaim… California        -626
##  2 Olive-Harvey College These ads ran without a disclaim… Illinois          -300
##  3 Chubb North America  These ads ran without a disclaim… Texas             -229
##  4 AARP Programs        These ads ran without a disclaim… Oklahoma          -228
##  5 First Republic Bank  These ads ran without a disclaim… California        -200
##  6 First Republic Bank  These ads ran without a disclaim… Massachuset…      -200
##  7 First Republic Bank  These ads ran without a disclaim… New Jersey        -200
##  8 First Republic Bank  These ads ran without a disclaim… New York          -200
##  9 First Republic Bank  These ads ran without a disclaim… Oregon            -200
## 10 Vacationvip.com      These ads ran without a disclaim… Arizona           -200
```

Addition-only paths

For comparison, here is an alternative path for combining the reports, which includes only summation operations.

```
## # A tibble: 7 x 4
##   report_date report_span_starts_on  span operation
##   <chr>       <date>                <int> <chr>
## 1 2020-04-17  2020-04-17                1 plus
## 2 2020-04-16  2020-04-16                1 plus
## 3 2020-04-15  2020-04-15                1 plus
## 4 2020-04-14  2020-04-14                1 plus
## 5 2020-04-13  2020-01-15               90 plus
## 6 2020-01-14  2020-01-08                7 plus
## 7 2020-01-07  2020-01-01                7 plus
```

It involves seven reports, and of them four are 1-day reports. (For comparison, the “plus-minus” path included only two 1-day reports.)

The table below shows, side by side, the spend amounts obtained using the “plus-only” path - column `s_p` , and the amount obtained using the “plus-minus” path - in column `s_pm` .

```
## # A tibble: 20 x 5
##    page_name      disclaimer             region             s_p    s_pm
##    <chr>          <chr>                  <chr>            <dbl>   <dbl>
##  1 Mike Bloomberg Mike Bloomberg 2020 Inc California     7076557 7076588
##  2 Mike Bloomberg Mike Bloomberg 2020 Inc Texas          5804710 5804727
##  3 Mike Bloomberg Mike Bloomberg 2020 Inc Florida        5339801 5339818
##  4 Mike Bloomberg Mike Bloomberg 2020 Inc Illinois       3220212 3220227
##  5 Mike Bloomberg Mike Bloomberg 2020 Inc Michigan       2931714 2931725
##  6 Mike Bloomberg Mike Bloomberg 2020 Inc Ohio           2793187 2793199
##  7 Mike Bloomberg Mike Bloomberg 2020 Inc North Carolina 2743139 2743144
##  8 Mike Bloomberg Mike Bloomberg 2020 Inc Virginia       2339973 2339981
##  9 Mike Bloomberg Mike Bloomberg 2020 Inc Pennsylvania   2339872 2339882
## 10 Mike Bloomberg Mike Bloomberg 2020 Inc Georgia        2298179 2298185
## 11 Tom Steyer     TOM STEYER 2020         California     2043556 2043649
## 12 Mike Bloomberg Mike Bloomberg 2020 Inc Massachusetts  2018089 2018096
## 13 Mike Bloomberg Mike Bloomberg 2020 Inc Washington     1738566 1738576
## 14 Mike Bloomberg Mike Bloomberg 2020 Inc Tennessee      1498338 1498343
## 15 Mike Bloomberg Mike Bloomberg 2020 Inc Arizona        1454704 1454711
## 16 Mike Bloomberg Mike Bloomberg 2020 Inc Colorado       1428649 1428655
## 17 Mike Bloomberg Mike Bloomberg 2020 Inc Minnesota      1392401 1392409
## 18 Tom Steyer     TOM STEYER 2020         South Carolina 1371596 1371631
## 19 Mike Bloomberg Mike Bloomberg 2020 Inc Missouri       1365768 1365775
## 20 Bernie Sanders BERNIE 2020             California     1324061 1324062
```

The agreement is very good.

Now, the table showing the entities where the “plus-minus” path produced negative numbers:

```
## # A tibble: 20 x 4
##    page_name                          region        s_p  s_pm
##    <chr>                              <chr>       <dbl> <dbl>
##  1 AARP Programs                      Oklahoma      100  -228
##  2 AARP Programs                      Texas         100  -187
##  3 AARP Programs                      Oregon        100  -115
##  4 The Late Show with Stephen Colbert California    100  -111
##  5 AARP Programs                      California    100  -106
##  6 Dopeaholics                        Alabama       300  -100
##  7 Dopeaholics                        Alaska        300  -100
##  8 Dopeaholics                        Arizona       300  -100
##  9 Dopeaholics                        Arkansas      300  -100
## 10 Dopeaholics                        California    300  -100
## 11 Dopeaholics                        Colorado      300  -100
## 12 Dopeaholics                        Connecticut   300  -100
## 13 Dopeaholics                        Delaware      300  -100
## 14 Dopeaholics                        Florida       300  -100
## 15 Dopeaholics                        Georgia       300  -100
## 16 Dopeaholics                        Hawaii        300  -100
## 17 Dopeaholics                        Idaho         300  -100
## 18 Dopeaholics                        Illinois      300  -100
## 19 Dopeaholics                        Indiana       300  -100
## 20 Dopeaholics                        Iowa          300  -100
```

Finally, a table showing the entities where both spend numbers were positive, but there was the largest discrepancy.

```
d_merged %>% mutate(d = amt_spent_p - amt_spent_pm, d_abs = abs(d)) %>%
  arrange(desc(d_abs)) %>%
  select(page_name, region, s_p = amt_spent_p, s_pm=amt_spent_pm, d) %>% slice(1:30)
```

```
## # A tibble: 30 x 5
##    page_name                            region        s_p  s_pm      d
##    <chr>                                <chr>       <dbl> <dbl>  <dbl>
##  1 NY State of Health                   New York      917 17832 -16915
##  2 New York City Department of Education New York    16152  3154  12998
##  3 Seniors Helping Seniors              Texas       17904  6281  11623
##  4 U.S. Census Bureau                   Connecticut  4501 15702 -11201
##  5 Edelson P.C.                         Georgia     14337  3182  11155
##  6 Chariot Energy                       Texas       12878  3254   9624
##  7 Veterans Advocates                   Texas       10993  1412   9581
##  8 U.S. Census Bureau                   Washington  15245  5926   9319
##  9 HealthInsurance.net                  Texas       12774  3622   9152
## 10 U.S. Census Bureau                   California  35044 27188   7856
## # … with 20 more rows
```

Conclusion

Due to variability in the quality of Facebook's reporting, we were facing the choice: go with the method that would minimize the round-off error - the “plus-minus” method, - but may end up with negative entries, or the method that may have have a higher round-off error but will have only positive numbers.

In the end, we felt that it is more important to avoid confusing the common users rather than worry about the round-off errors. In addition, it appears that Facebook is more likely to have errors for small advertisers, but the numbers for large advertisers converge, regarding the method.

Therefore, our final choice is the “plus-only” method.