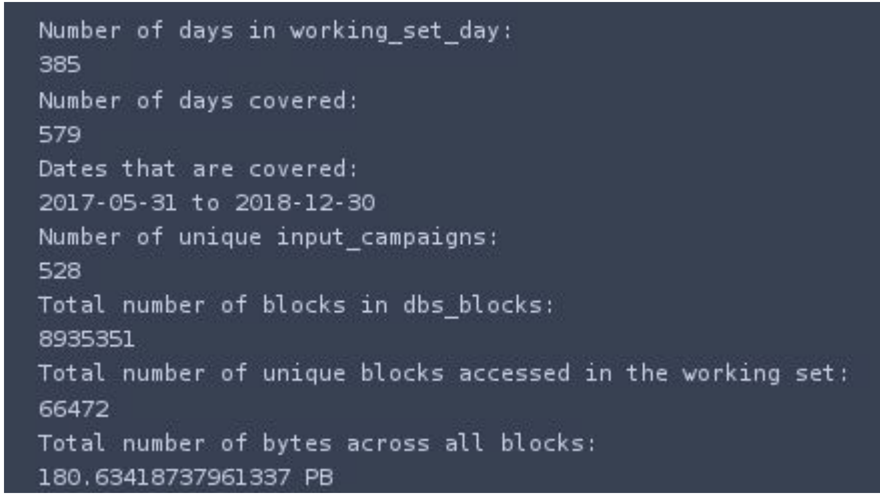
6-04-19 Weekly Report

Previous Goals

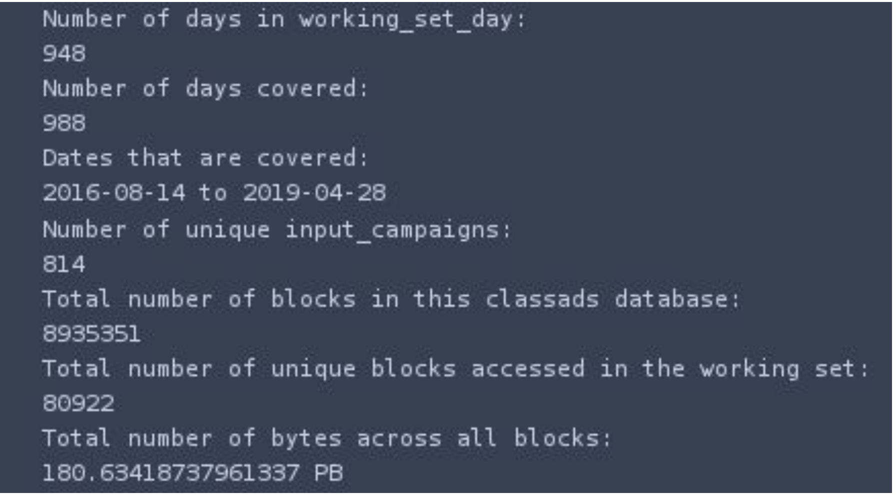
- ~~Do a characterization of Diego’s newer working set~~
- ~~Get bytes and gained/dropped block code to work~~
- ~~Drop everything that is not crab_job~~
- ~~If in the newer working set, the spike still exists, investigate spike (check to see if it is a bug) (Investigating blocks accessed around January 2018 individually and compare to January 2019)~~
- Find out what datasets the blocks in the spike corresponds to
- Extract 10 rows of these blocks so that Frank can discern which datasets correspond to those blocks so that you don’t have to guess

Characterization of Diego’s Working Set



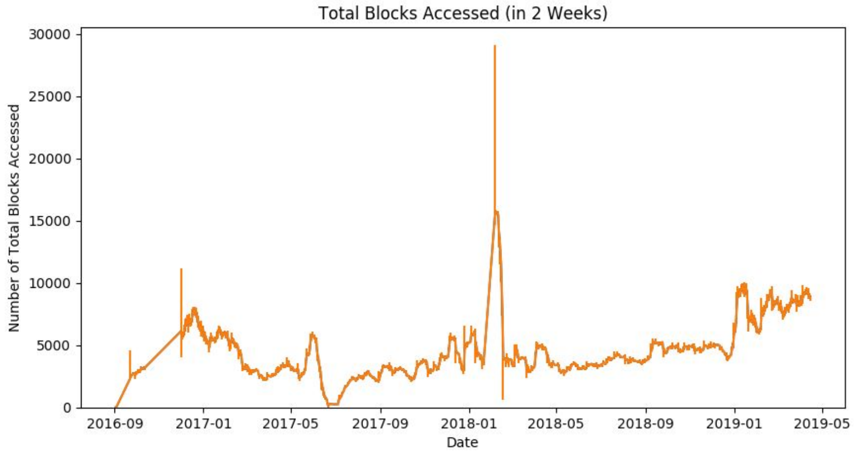
```
Number of days in working_set_day:
385
Number of days covered:
579
Dates that are covered:
2017-05-31 to 2018-12-30
Number of unique input_campaigns:
528
Total number of blocks in dbs_blocks:
8935351
Total number of unique blocks accessed in the working set:
66472
Total number of bytes across all blocks:
180.63418737961337 PB
```

Characterization of Old (2nd) Working Set

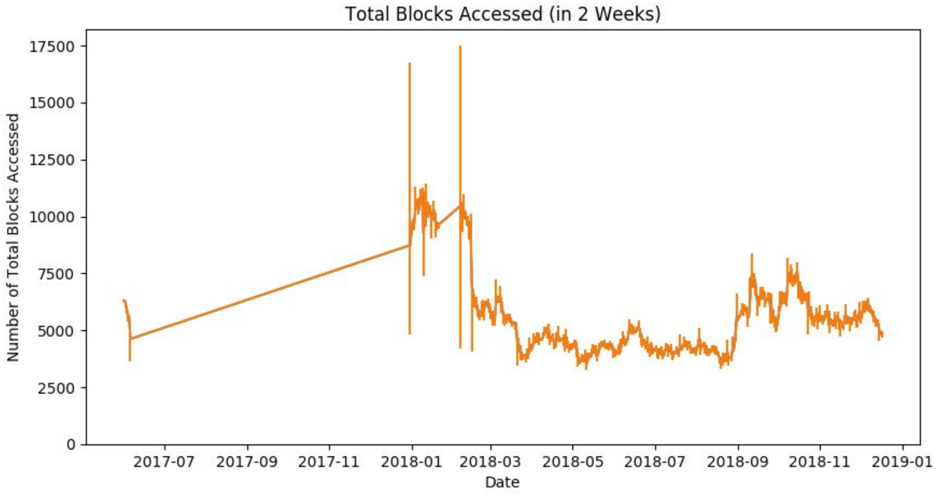


```
Number of days in working_set_day:
948
Number of days covered:
988
Dates that are covered:
2016-08-14 to 2019-04-28
Number of unique input_campaigns:
814
Total number of blocks in this classads database:
8935351
Total number of unique blocks accessed in the working set:
80922
Total number of bytes across all blocks:
180.63418737961337 PB
```

New Dataset

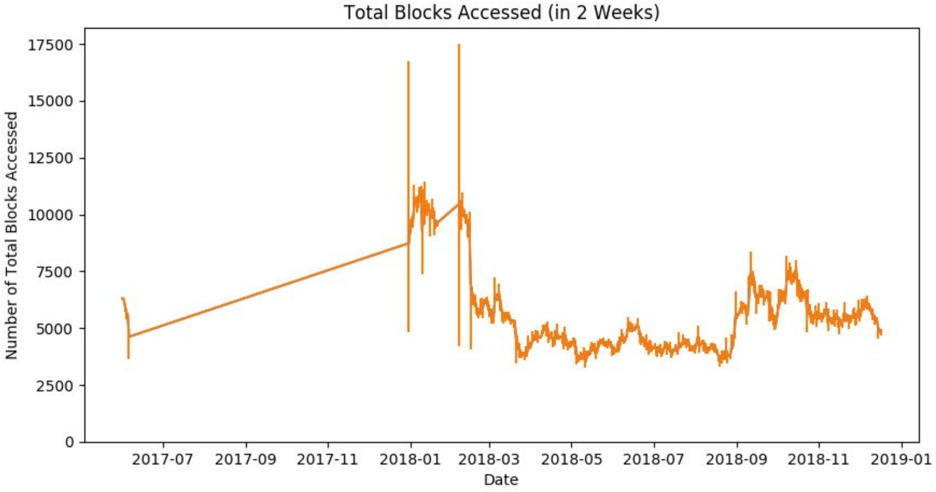


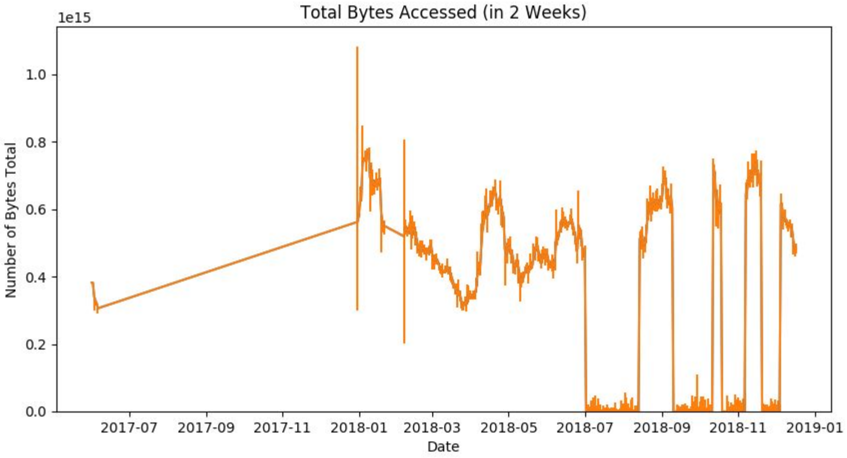
The first plot shows the total blocks accessed (in a 2 week period) with blocks dropped/gained as error bars for the 2nd dataset (plot from 3 weeks ago). This is the one with a noticeable giant spike after January 2018.



The 2nd plot shows the same for Diego’s new dataset (but with data only for 2018). A spike does exist, but not at the same magnitude as the previous plot. However, it appears that the reason for this spike is that there is a gap in data accesses before the spike. Because this spike is a result of a gap in data as opposed to any particular datasets, I won’t investigate which datasets correspond to the spike.

New Dataset (Bytes)





There have been some issues in summing up the bytes. For some select time periods, there were errors when trying to access the bytes of certain blocks. For now, I’ve just dropped that data, but the rest remains intact.

Extensions

- Talk to Diego about the gap/inspect months for gaps
- Debug byte summing code