GHTorrent: Github's Data from a Firehouse

The authors reverse engineered the structure of github and wrote crawlers to extract information about repositories. The result is distributed through Torrenting which is a peer-to-peer sharing method that's widely used. The dataset is constructed such that updates can be handled by a batch based data structure using MongoDB which ease the process of updating data. Some simple analysis of the result is also shown to illustrate the usefulness of this dataset.

Things I would like to see discussed:

1. More information on how Github's API restriction can be bypassed and why Github does not provided such data by themselves.
2. Generizability to other services such as GitLab due to recent drama on Microsoft acquiring Github
3. REST and what it does and how it affects the process of retrieving data.