Started with a table of the 200k repos which are forks and have the most Push events (minimum 4).

Made a table concerning each fork/parent pair on which we have full data (178,039) of the original list of 200k. This wide table allows us to read into the history of a relationship between a fork and its parent conveniently.

**Looking at the first row of the augmented data-set:**

Parent = “shadow386/CanaryMod” Fork = “FallenMoonNetwork/CanaryMod”

The parent repository was created on 10th October 2011 and was active until 28th March 2012 – seeing 72 PushEvents in this time by 3 different users (42 coming from the repo's owner). By the time timeline begins this repo has 44 watchers and 21 forks, these figures grow only slightly by the time the last push event is observed in March 2012 (to 48 and 23 respectively). During the timeline data the parent saw 2 incoming pull requests from 2 distinct repos and made no outgoing pull requests.

On 27th March 2012 the fork repo “FallenMoonNetwork/CanaryMod” was created, so one day before the last push on the parent repo. The fork's last recorded push was on 4th June 2013, so its likely still active – during this time it has received 681 pushes from 10 different users. Its number of watchers has grown from 1 to 26, and it has been forked 21 times. It has received 33 pull requests from 14 different repos (occurring between 4th April 2012 and 31st March 2013), and has made 2 pull requests with itself as the head repo. None of the pull requests for these repos are between them, so thus far there is no sign of a relationship between the repos (other than the initial fork) - and it looks like a case of the fork repo suplanting its parent.

However, if we look at the users who have pushed to the fork and its parent repo (the most time-consuming column of this table to assemble) it turns out that the 3 pushers on the parent are all among the 8 users who have pushed to the fork. So, the interpretation of this fork/parent relationship changes considerably, and it begins to look more like a simple re-naming procedure (in the timeline what we did at the beginning with our repo(s) would look very similar).

This suggests one follow-up which could fairly easily be tracked down with a visit back to bigquery: who are the additional five users who have joined the project since it was forked, there is still a possibility that this was not re-naming but some sort of merger or takeover. We could look at whether these new users made pull requests on the original parent repo or for other signs of a link to the project before it was forked. We could also look at raw activity on the fork repo – are the original 3 users still the most active on the new forked repo, or are they taking a back seat? Simple push event counts could have something to say about that, it does involve looking more at users but I think cases like this demonstrate that sometimes we can't get the full picture on the relationship between repos without looking at the people involved.

**Getting a feel for the data-set:**

29,693 repos where fork has more pushes than parent.

For 24,746 of these there is no overlap between fork/parent pushers.

Setting some (stringent) criteria for what it would look like for a fork to supplant its activity... More pushes, more watchers and forks growth during timeline period, more pull requests received (and also no common pushers, to rule out re-naming or somesuch).

There are 5,662 repos which meet these criteria – 6929 if we include repos that share common pusher(s). However, in a lot of cases these pairs are fairly small-scale (2,801 of the forks have 10 or less push events).

**Regarding Pull Requests between the fork and parent repo:**

In the full set there are only 3,763 relationships where a pull request has been made between fork/parent repo with the parent as the base. There are 5,382 relationships where a pull request has been made with the fork as base. These numbers both seem unusually low, and its perhaps strange that there are fewer relationships having a pull request with parent as base.

**Re-naming 'Relationships'**

Where all of the pushers to the base repo become pushers on the forked repo this could be something like a re-naming operation, if the fork and parent share a common and solitary pusher this is almost certainly the case.

There are 3,637 relationships where a single user is the sole pusher to both fork and parent repository.

There are 5,000 repos in total where all parent pushers became pushers on the fork.

**Fork/Parent overlap**

Is there an overlap of activity between fork and parent repo? Effectively, are the parent repos in this data-set still active when the forks are created? Comparing last push on the parent to creation data of the fork. 15,129 pairs where the parent sees no more pushes after the fork is created. There are a further 4,968 pairs where the parent's last push was recorded within 24 hours of the fork's creation.