1 Introduction

2 Research Question

3 Related Work

4 Study Setting

4.1 Description of data-set

In this project we have used peer-review data from Chromium project as data for our analysis. This data was obtained from the Chromium database which is available on the internet for public use. The raw data from the Chromium database contained four tables persons, review, comment and approval. As the names suggest, the table persons contained information on the various people (i.e developers) who contributed. It has two columns owneremail and owner both of which contain unique values of email ids and user names. The second table, review contains contains information on each review. It's columns are:

1.issue: This column contains a unique issue number by which each review is identified. 2.owner: This columnc holds the user names of the developers who have created the review. 3.description: This column contains a brief description of the review 4.subject: This column tells us the matter addressed by the review 5.created: Date of creation of the review 6.modified: Date of last modification of the review.

The third table comment contain data on the comments made by the developers on different reviews. It conatins the following columns:

1.id: This is an unique id assigned to each comment 2.issue: This column contains a unique id of the review that was commented on. 2.sender: This columnc holds the user names of the developers who have made that particular comment. 3.recipients: Contains the name of the recipient. 4.text: This column contains the comment 5.disapproval: Holds a boolean value which indicates whether the comment has been disapproved or not 6.date: Date when the comment was made 7.approval: Holds a boolean value which indicates whether the comment has been approved or not. The fourth table is called approval. It has four columns namely : 1.issue: This is the issue number of the review by which it can be identified 2.owner: Contains the name of the owner 3.closed: Date of closing of the review 4.commit: Contains a boolean value which indicates whether a particular review has been committed or not.

4.2 Filtering and pre-processing data

The raw data obtained from the Chromium database needed to to filtered to extract the relavent information for the analysis. For our analysis we needed to create a social network of developers where any two developer are considered

to be related to each other if they have commented on a common review. The number of such reviews thus represent the edge weights in the network.

As stated above, the raw data from the Chromium database contained four tables persons, review, comment and approval. However, we do not require all of this data to perform the analysis. This calls for a method for flitering and pre-processing of the data.The required data was obtained by querying in mysql. The database was filterd to extract data on the number of comments per review, no if people who have commented on a review, no of approvals on the review, date of creation, date of modification , no of days between date of creating and date of modification,total no of reviews owner by each person who owns a review. The following code was used to extract data from the raw data

select * from (select t.comments, t.NoOfCommenters, t.approval, r.issue, r.created, r.modified, datediff((cast(r.modified as date)),(cast(r.created as date))) noofdays from review r left join (select issue, count(issue) comments, count(distinct sender) NoOfCommenters, count(case approval when 'true' then 1 else NULL end) approval from comment group by issue) t on t.issue=r.issue order by r.issue asc) r1 natural join (select t.owners, r.issue from review r left join (select owner, count(owner)owners from review group by owner)t on t.owner=r.owner order by r.issue asc) r2;

4.3 Network generation

Using the filtered data a of developers network was generated over a span of 50 cumulative time intervals . Each developer represents nodes in the newtwork. Any two developer are considered to be related to each other if they have commented on a common review. The number of such reviews thus represent the edge weights in the network. The edge weights were caluculated using the pandas library and the network was generated using the igraph library in python .

4.4 Calculation of network parameters

5 Results

6 Discussion

7 Threats to validity and future work

8 Summary and Conclusions