Empirical Software Engineering using Ultra Large Repositories

Mei Nagappan
Agenda

• Part 1 - Introduction
  • Course Overview Objectives
  • Student introduction and expectations
  • Syllabus and Project
• Part 2 - Example of an ESEULR study
• Part 3 - Intro to stats.
Typical ESE vs ESE in ULR
What can we learn about SE from these Ultra Large Repositories?
Challenges
Part 2
Example Study

How do ratings evolve?
Are Most Apps Great?

Fig. 1: Distribution comparison of the global rating of the last version of all analyzed mobile apps in - and mobile apps with at least two versions and at least - raters per version. The X axis is in terms of probability. 

Fig. 2: Natural logarithm of the minimum and maximum number of raters across all versions of each app with at least two versions in - . The dashed line shows the threshold for the minimum number of raters per version used to filter the data. 

Fig. 3: Global ratings vs. the natural logarithm of the number of raters for the last version in - of all mobile apps with at least two versions in - .

Fig. 4: 5th/50th/75th percentiles per app category of the number of raters in the last version of the mobile apps with at least two versions in - and - raters per version. 

Many apps contain biased ratings because of a low number of versions and/or raters. The left-hand side of Figure 1 shows the distribution of the global rating of the last version of all apps in our data. We note that a large proportion of apps has a rating close to 5. With ,83 % of the apps having a rating of exactly 5, the median rating is % and we notice local peaks for ratings % and %. However, this distribution is biased for a number of reasons. First, not every app attracts the same level of user feedback in the form of ratings. Figure 2 shows the minimum and...
Lots of Apps with very few Ratings

128K+

10K+

Counts

Fig. (, Distribution comparison of the global:rating of the last version of all analyzed mobile apps in and mobile apps with at least two versions and at least raters per version, The X axis is in terms of probability.

Fig. ), Natural logarithm of the minimum and maximum number of raters across all versions of each app with at least two versions in and the dashed line shows the threshold for the minimum number of raters per version used to filter the data, Note that the logarithm of is plotted as .

While during the time of data collection Google Play only showed the resulting global:rating, it now shows the breakdown of this rating across to 5 stars, ratings lower than star are not allowed, We used this breakdown to understand special cases and outliers.

http://goo.gl/t6d8
Most apps are Average
More Raters => Steady Ratings

Figure 03 Scatterplot of global rating increase between the first and last version of apps with at least 7 versions in 7566 versus the total number of raters up until the first version in 7566.

Counts

Fig 3.3 Scatterplot of increase in local rating for all versions of the apps with at least 7 versions in 7566 and 65 raters per version versus the corresponding increase in global rating.

Findings

Once a significant number of people have rated an app, the app's global rating is resilient to fluctuations in rating.

Figure 0 plots for each app in our filtered data set the increase in global rating between their first and last version in 7566 versus the total number of people who had rated a version up until the first version in 7566. Given that the X axis is in logarithmic scale, the plot clearly shows that the more people have voted for an app throughout its lifetime, the more resilient the app becomes to changes in its rating. In other words, large apps technically have some leeway to experiment in a new version without their rating being affected negatively.

We looked at the history of the four apps with the largest number of raters to see how their global and local rating has fluctuated from 7566 to the last version today in 7567.

Facebook
- Raters initially: 1.18e+05
- Global rating: 8.3
- Current rating: 8.31

Google Maps
- Raters initially: 8.17e+05
- Global rating: 9.30
- Current rating: 9.39

Angry Birds
- Raters initially: 1.90e+05
- Global rating: 9.30
- Current rating: 9.31

Kakao Talk
- Raters initially: 7.77e+05
- Global rating: 9.30
- Current rating: 9.30

These results suggest that the real quality of an app does not really matter once a massive number of users favorably rated the app, since the advertised user ratings remain the same. However, as mentioned in RQ6, the rating advertised by Google Play is basically a global rating, i.e., an average across all raters' ratings across time. While this global rating might be resilient, the local rating of a specific app version provides a more instantaneous and hence accurate view of the perceived quality of a particular app.

Some increases in local rating can cause global rating increases larger than 0.5.
More Raters => Steady Ratings

Figure 03 Scatterplot of global rating increase between the first and last version of apps with at least 7 versions in 7566 versus the total number of raters up until the first version in 7566.

Figure 03.3 Scatterplot of increase in local rating for all versions of the apps with at least 7 versions in 7566 and 65 raters per version versus the corresponding increase in global rating. Furthermore, to calculate probabilities instead of density functions, we discretized the local rating into "star ratings" by rounding off. For example, a local rating of 930 becomes 01 whereas 938 becomes 93. Such a star rating is akin to how users input their ratings for an app.

Findings:

Once a significant number of people have rated an app, the app's global rating is resilient to fluctuations in rating. Figure 0 plots for each app in our filtered data set the increase in global rating between their first and last version in 7566 versus the total number of people who had rated a version up until the first version in 7566. Given that the X axis is in logarithmic scale, the plot clearly shows that the more people have voted for an app throughout its lifetime, the more resilient the app becomes to changes in its rating. In other words, large apps technically have some leeway to experiment in a new version without their rating being affected negatively.

We looked at the history of the four apps with the largest number of raters to see how their global and local rating has fluctuated from 7566 to the last version today in 7567. Facebook initially had 9188 global raters, now 83.1 dropped slightly from a global rating of 83.1 to 83.3 while the local rating increased once from 83.8 to 83.3. Google Maps initially had 715.5 raters, now 9391 dropped slightly from 930 to 9391 while the local rating decreased once by 539, and increased twice by 5381 ending at 93.3. Angry Birds initially had 617,.1 raters, slightly increased from 930 to 93.1 while locally the rating increased from 930 to 93:3. Finally, KakaoTalk initially had 8.19 raters, globally stayed on 930 while locally the rating dropped once by 6309 and increased once by 6399. This last case is the most representative example of an app with stable global rating, however having highly fluctuating local ratings.

These results suggest that the real quality of an app does not really matter once a massive number of users favorably rated the app, since the advertised user ratings remain the same. However, as mentioned in RQ6, the rating advertised by Google Play is basically a global rating, i.e., an average across all raters' ratings across time. While this global rating might be resilient, the local rating of a specific app version provides a more instantaneous and hence accurate view of the perceived quality of a particular app.

Some increases in local rating can cause global rating increases larger than 0.5. Figure . plots for each successive pair of app versions the increase in local rating versus the increase in global rating. We can see three phenomena. First, the majority of rating changes float around 5 for both local and global rating (dark black cells), i.e., this is the case where no rating change happens. Second, a long vertical stretch
Low Local Rating $\Rightarrow$ Stable

More than 1 star drop $\Rightarrow$ Unrecoverable

High Local Rating $\Rightarrow$ Unstable

Findings

Once a significant number of people have rated an app, the app's global rating is resilient to fluctuations in rating. Figure 0 plots for each app in our filtered data set the increase in global rating between their first and last version in 2016 versus the total number of people who had rated a version up until the first version in 20163. Given that the X axis is in logarithmic scale, the plot clearly shows that the more people have voted for an app throughout its lifetime, the more resilient the app becomes to changes in its rating. In other words, large apps technically have some leeway to experiment in a new version without their rating being affected negatively.

We looked at the history of the four apps with the largest number of raters to see how their global and local rating has fluctuated from 2016 to the last version today in 2017. Facebook initially rated 8.3, now rated 8.3.1, dropped slightly from a global rating of 8.3 to 8.3.1 while the local rating increased once from 8.3 to 8.3.3. GoogleMaps initially rated 9.3 to 9.391, while the local rating decreased once by 539, and increased twice by 538, ending at 9.3.3. AngryBirds initially rated 9.3 to 9.3.1, while locally the rating increased from 9.3 to 9.3.3. Finally, KakaoTalk initially rated 9.30, globally stayed on 9.30, while locally the rating dropped once by 9.309 and increased once by 9.399. This last case is the most representative example of an app with stable global rating, however having highly fluctuating local ratings.

These results suggest that the real quality of an app does not really matter once a massive number of users favorably rated the app, since the advertised user ratings remain the same. However, as mentioned in RQ6, the rating advertised by Google Play is basically a global rating, i.e., an average across all raters' ratings across time. While this global rating might be resilient, the local rating of a specific app version provides a more instantaneous and hence accurate view of the perceived quality of a particular app.

Some increases in local rating can cause global rating increases larger than 0.5.
Part 3
Stats

• Mean/Median/Mode
• Histogram
• Correlation
• Hypothesis testing
• Regression