Learning a Codebase

Derek Somerville

November 17, 2025

Contents

1 Repository

1.1 Glossary - Summary

- The founder developer starts in the first six months of a project.
- The late joiner developers start after six months.
- Sustained developers make 50 or more commits and commit for 250 days or more.
- Transient developers have fewer commits or commit for a shorter period.

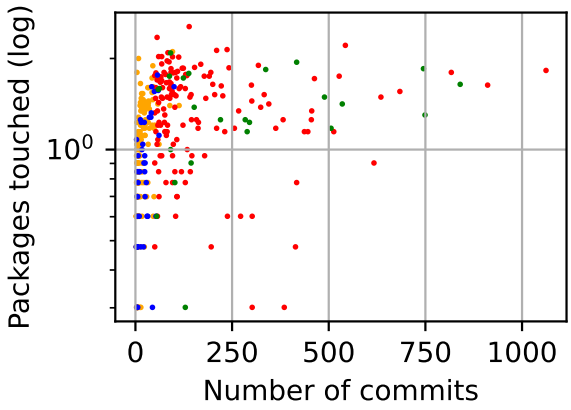
1.2 Repository Summary Table

Table 1: Summary of fifteen open-source repositories identified from GitHub that have at least 1000 pull requests and at least three sustained late joiner developers. Sustained later joiner developers joining the project after six months and contributed at least 50 commits over a period of 250 days or more. Developers with fewer than three commits are excluded. Please note that five developers each worked on two repositories.

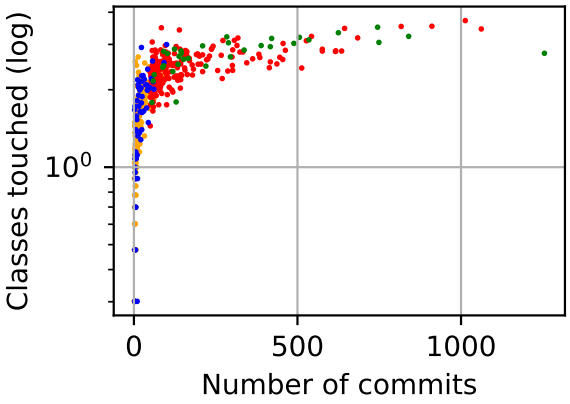
ID	Repo Name	Founder Transient Contributor	Founder Sustained Contributor	Joiner Transient Contributor	Joiner Sustained Contributor	Commit	Start	End
3	activiti	11	6	70	5	2329	2010-Jun-18	On going
5	airbyte-platform	8	1	93	6	1712	2020-Aug-04	2023-Sep-27
7	ambari	4	0	76	15	4125	2011-Sep-22	2023-Nov-18
10	automq	2	0	65	3	616	2011-Sep-07	2017-Apr-21
14	buck	13	6	204	21	6316	2013-Mar-21	2021-May-17
15	camel	5	1	328	13	5441	2007-Mar-19	On going
18	checkstyle	1	1	112	9	2803	2001-Jun-22	On going
20	cxf	12	3	44	3	1687	2008-Apr-29	On going
21	intellij-community	3	7	215	84	28741	2004-Nov-11	2018-Apr-18
26	guava	1	0	60	3	955	2009-Sep-15	2024-Jan-23
28	jenkins	2	1	201	4	3594	2006-Nov-05	On going
33	openmrs-core	1	0	162	3	1466	2006-May-03	On going
36	presto	1	3	244	18	4867	2012-Aug-09	On going
37	quarkus	4	5	167	8	2682	2018-Jun-22	On going
39	selenium	6	0	79	8	1718	2004-Nov-03	On going
	Total	74	34	2120	203	69052		

2 Sample scatter touched against commit -

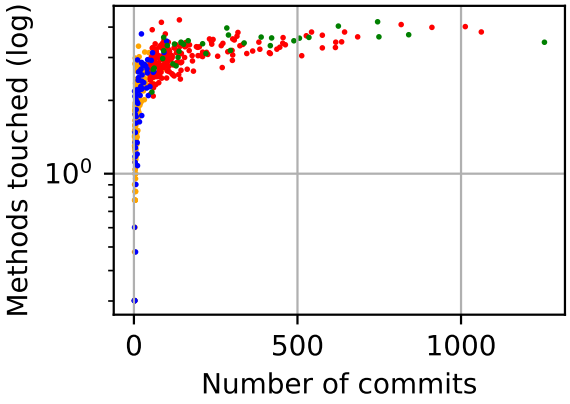
A scatter plot of touched components by number of commits. Developers with fewer than three commits are excluded.



(a) Packages



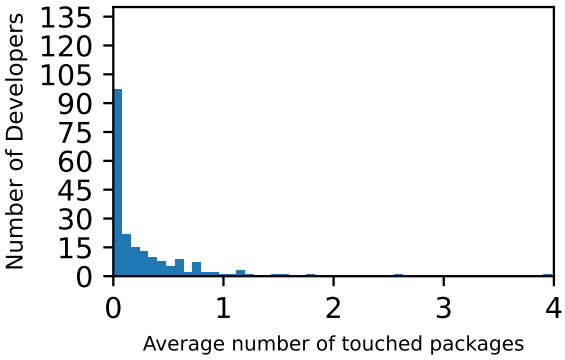
(b) Classes



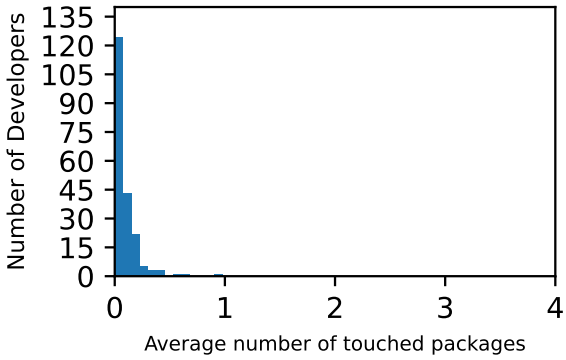
(c) Methods

Figure 1: Scatter plots of the log total number of components touched (y-axis) against the number of commits (x-axis) made by samples of developers capped at 203 from four categories: Transient Founder (Blue, 68), : Sustained Founder (Green, 34), : Transient Later Joiner (Orange, 203), : Sustained Later Joiner (Red, 203).

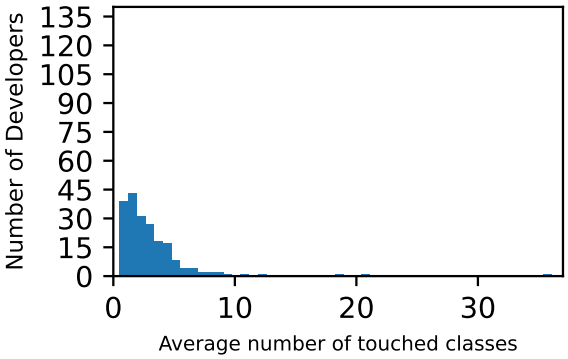
3 Repository average touched - By commit and by day



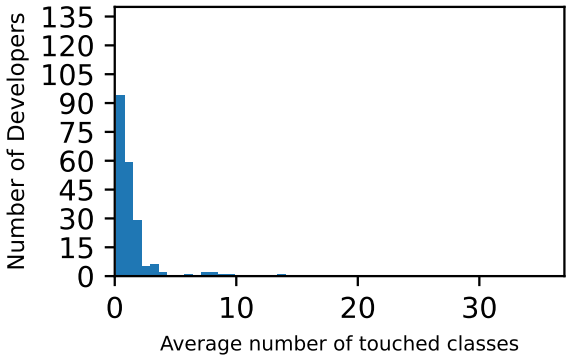
(a) Commit new packages.



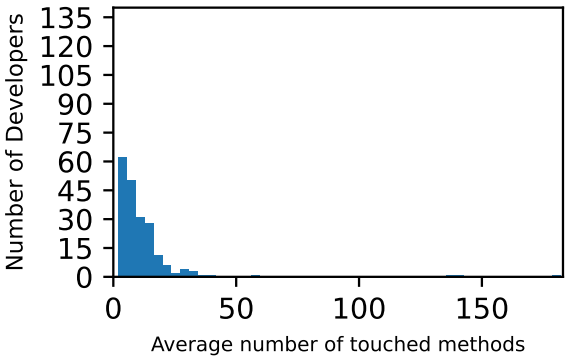
(b) Daily new packages.



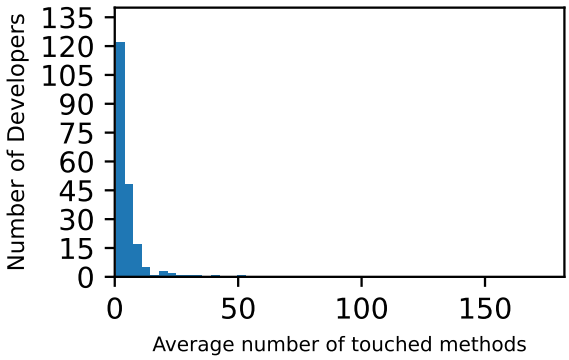
(c) Commit new classes.



(d) Daily new classes.



(e) Commit new methods.



(f) Daily new methods.

Figure 2: Average number of new components touched per day and per commit (x-axis) for 203 sustained later joiner developers (y-axis) from fifteen (15) projects sampled from GitHub.

3.1 Developer commit - Developer Commits

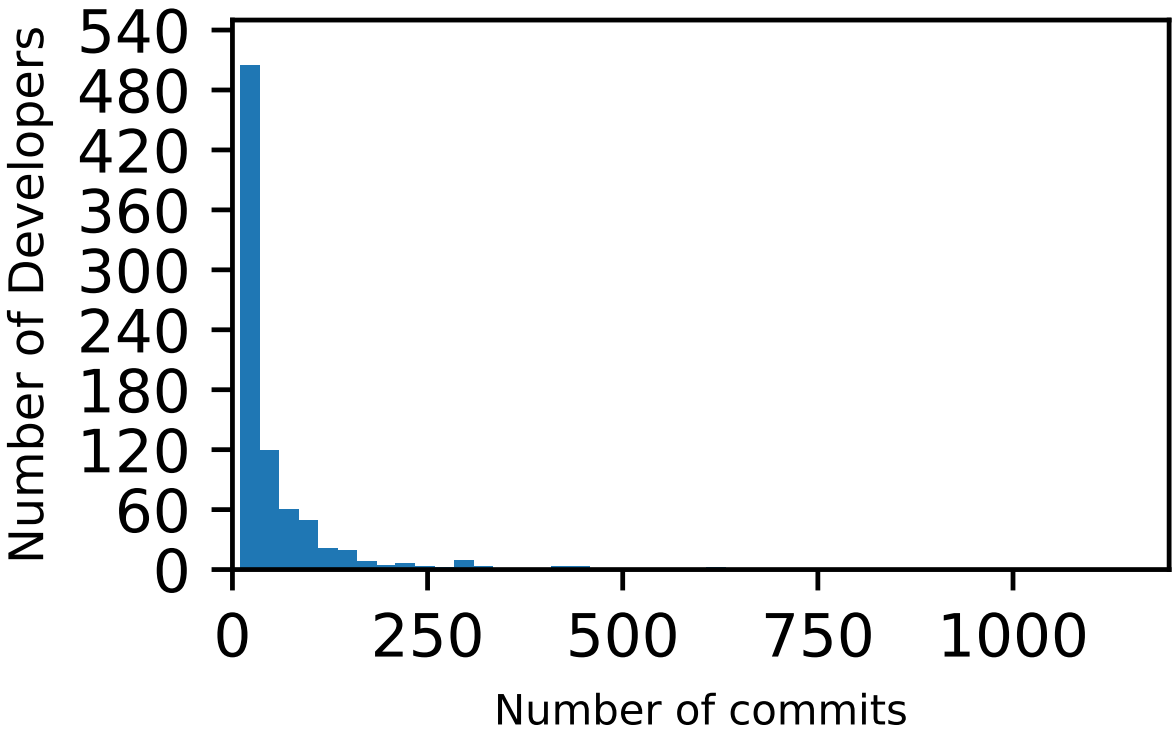
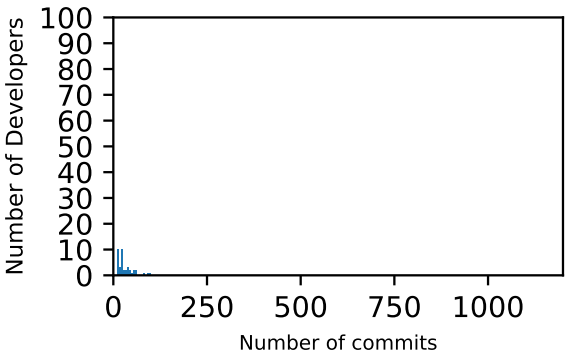
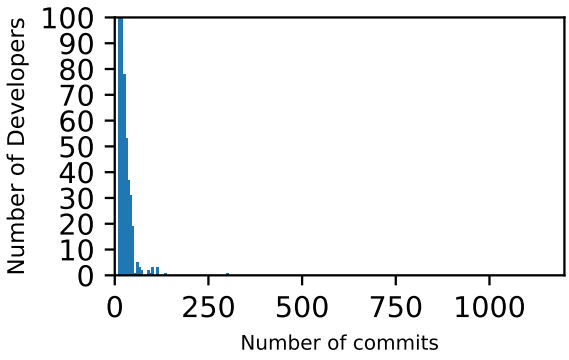


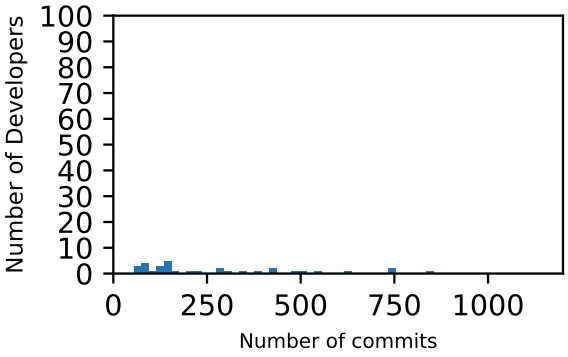
Figure 3: All 857



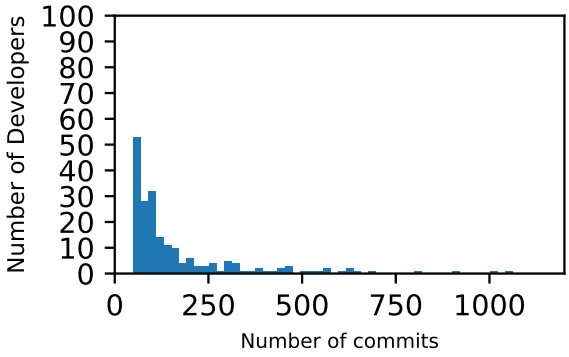
(a) transient founder 40



(b) transient later joiner 580



(c) sustained founder 34

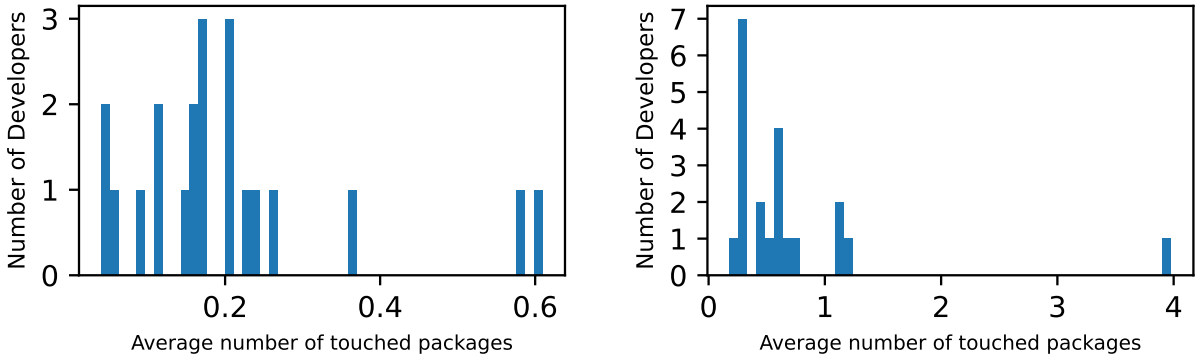


(d) sustained later joiner 203

Figure 4: Histomagram of number of commits made by four categories of developers, excluding developers with less than ten (10)

4 Repository: 14

4.1 Repository histogram commit - 14 For packages with total



(a) By Day.

(b) By Commit.

Figure 5: Histogram of new packages touched (x-axis) against number of developers (y-axis). Graphs for 21 sustained later joiner developers joining the project after six months and contributed at least 50 commits over a period of 250 days or more.

4.2 Time series developer - 14 For packages touched for each period

A time series of packages touched on average each month.

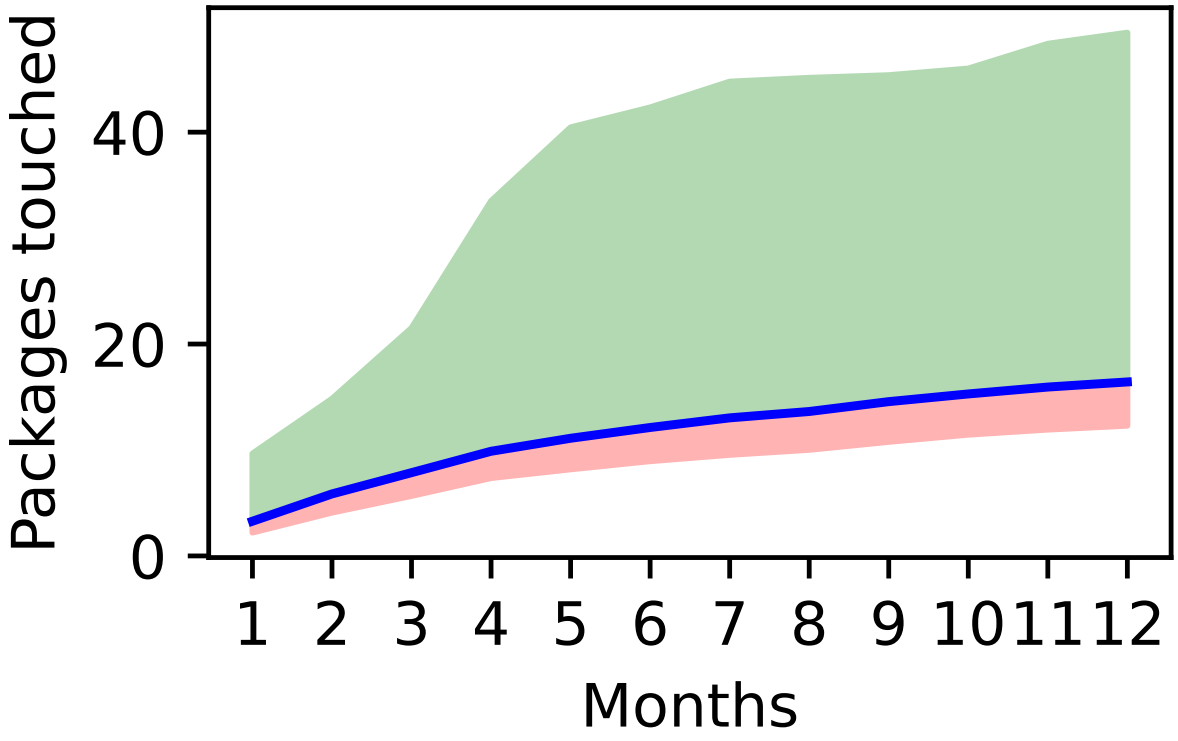
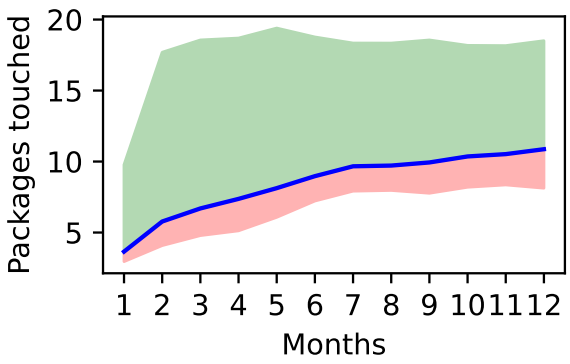
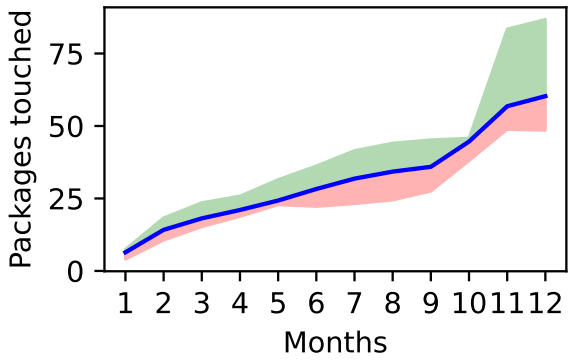


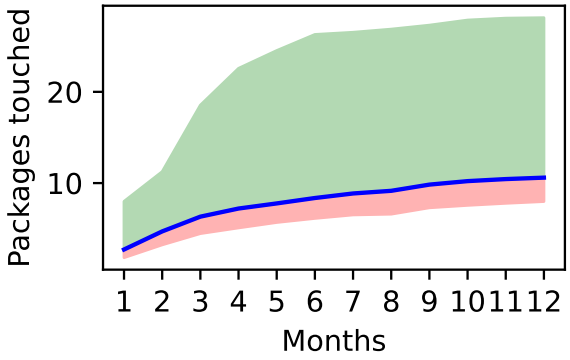
Figure 6: All developers (244) showing packages touched



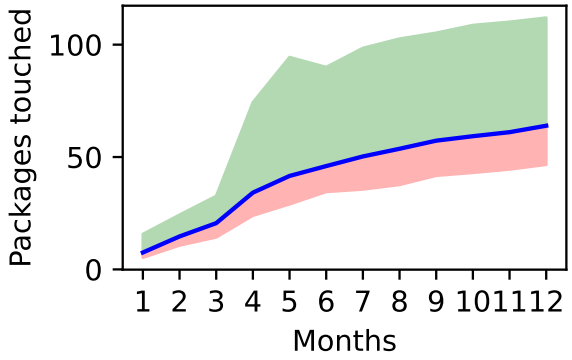
(a) Transient founder developers (13) showing packages touched



(b) Sustained founder developers (6) showing packages touched



(c) Transient later joiner developers (204) showing packages touched



(d) Sustained later joiner developers (21) showing packages touched

Figure 7: A time series of the number of month (x-axis) against the average (mean) total packages touched (y-axis), with positive (orange) and negative (red) filled standard deviation.

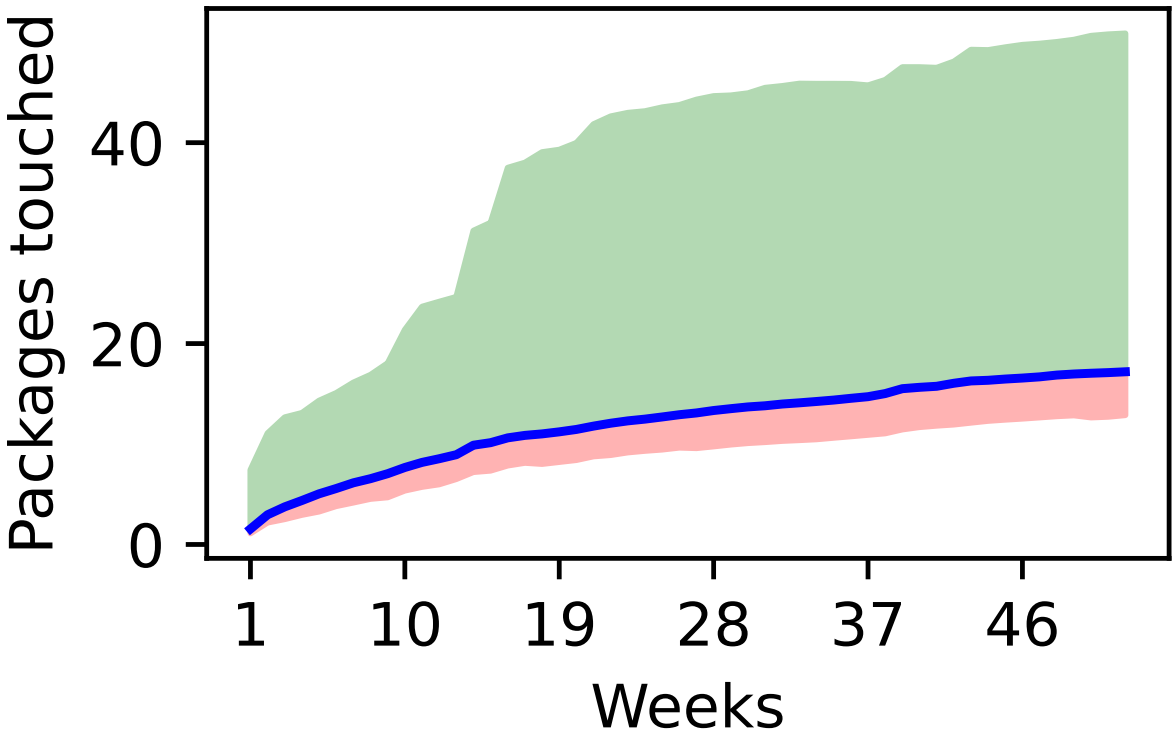
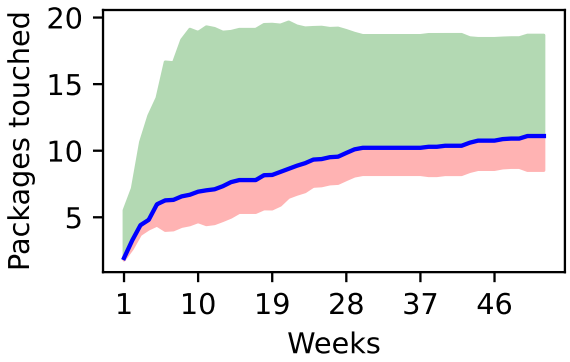
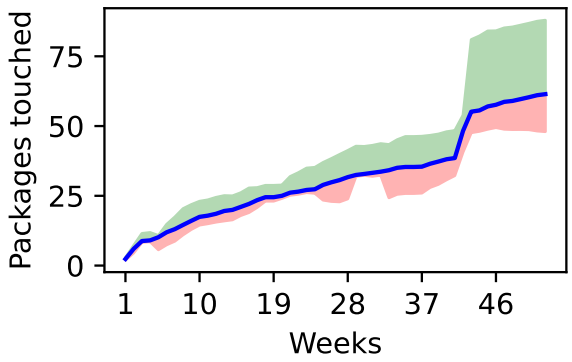


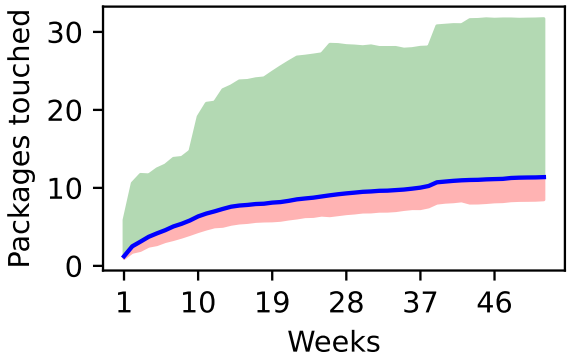
Figure 8: All developers (244) showing packages touched



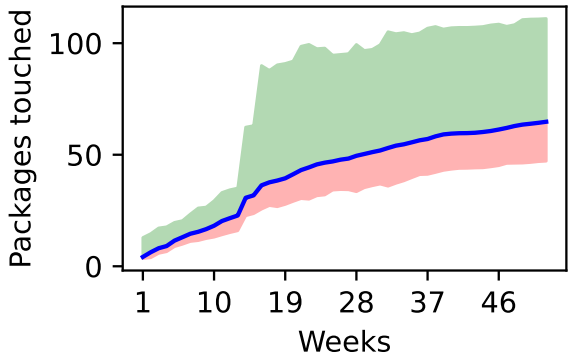
(a) Transient founder developers (13) showing packages touched



(b) Sustained founder developers (6) showing packages touched



(c) Transient later joiner developers (204) showing packages touched



(d) Sustained later joiner developers (21) showing packages touched

Figure 9: A time series of the number of week (x-axis) against the average (mean) total packages touched (y-axis), with positive (orange) and negative (red) filled standard deviation.

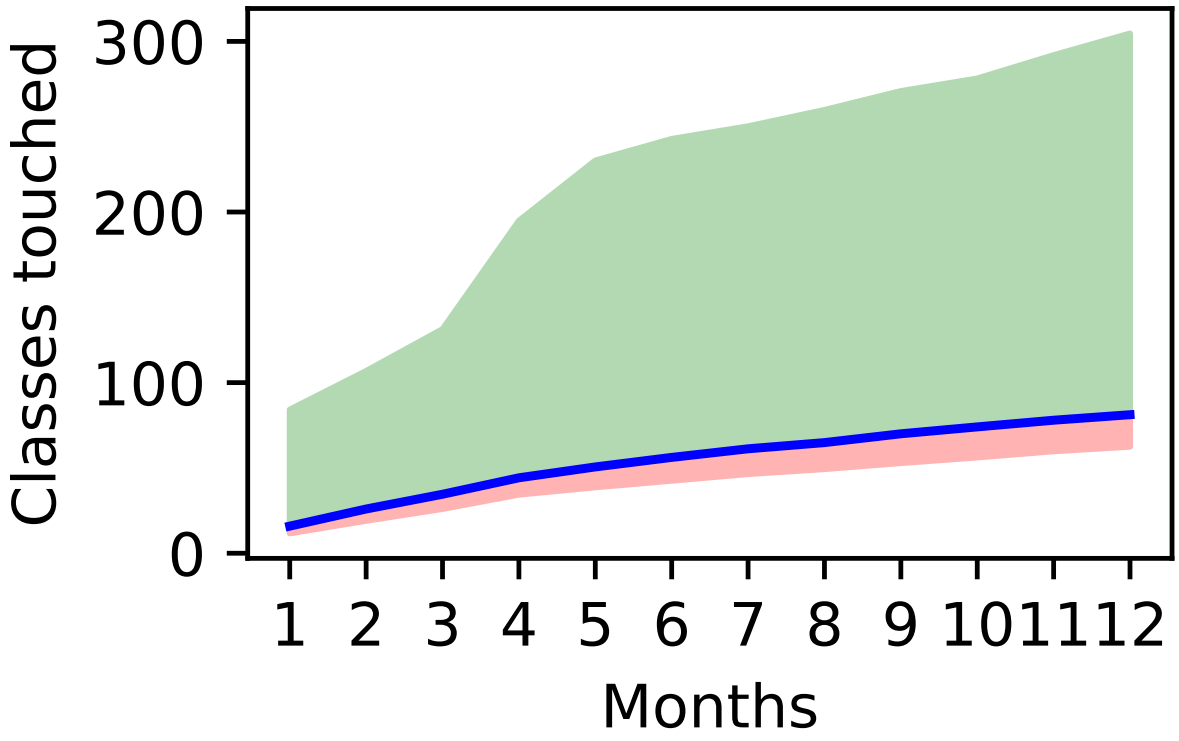
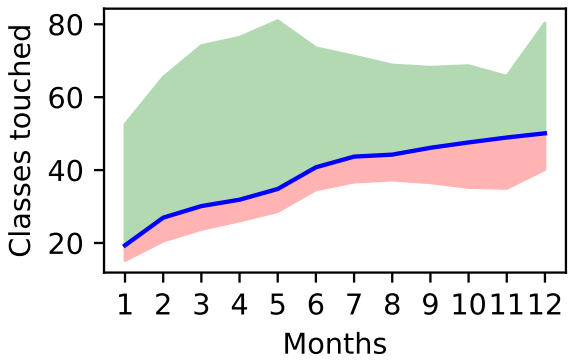
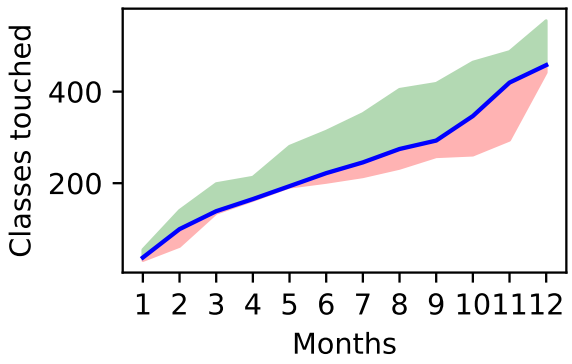


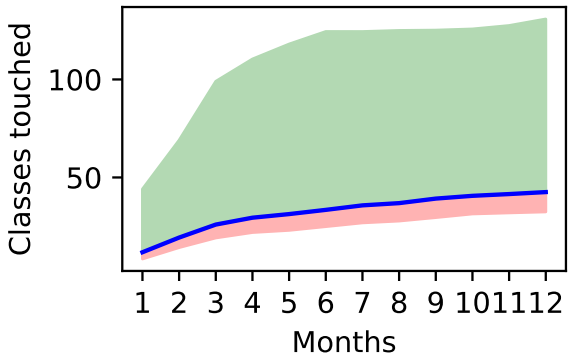
Figure 10: All developers (244) showing classes touched



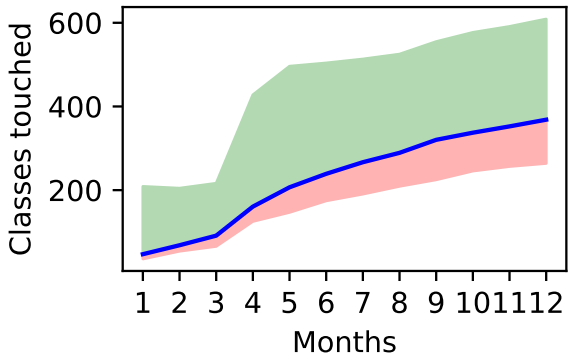
(a) Transient founder developers (13) showing classes touched



(b) Sustained founder developers (6) showing classes touched



(c) Transient later joiner developers (204) showing classes touched



(d) Sustained later joiner developers (21) showing classes touched

Figure 11: A time series of the number of month (x-axis) against the average (mean) total classes touched (y-axis), with positive (orange) and negative (red) filled standard deviation.

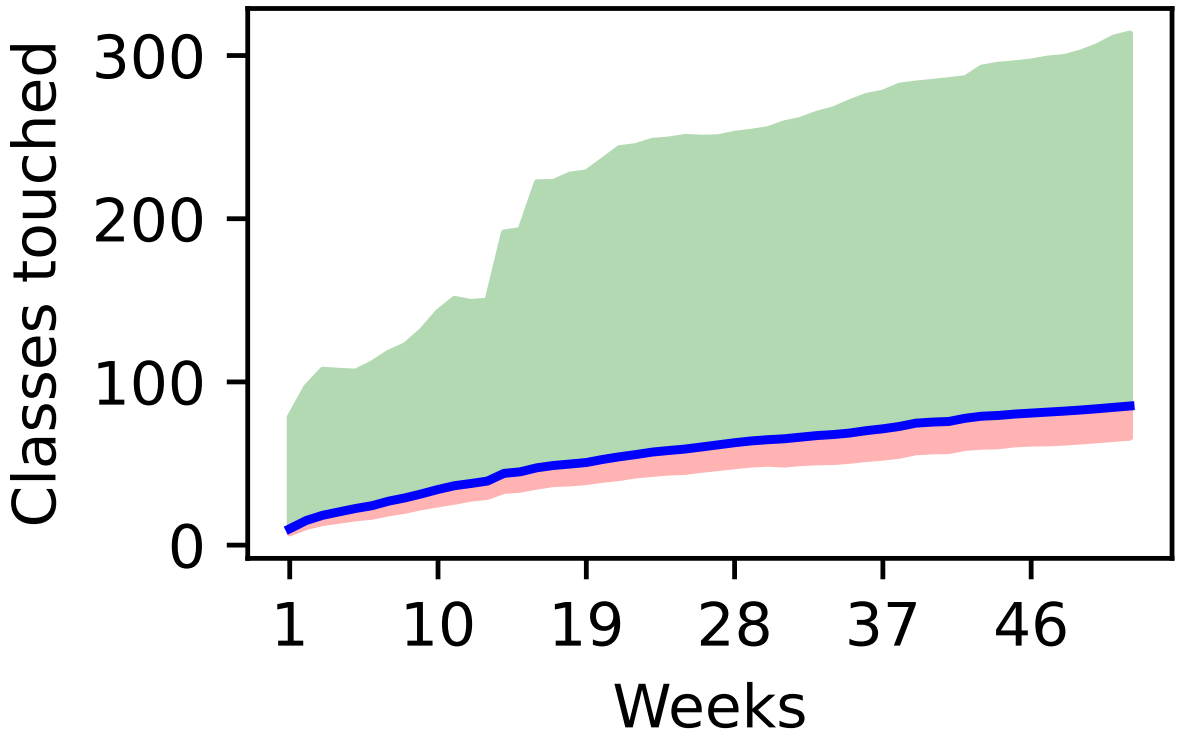
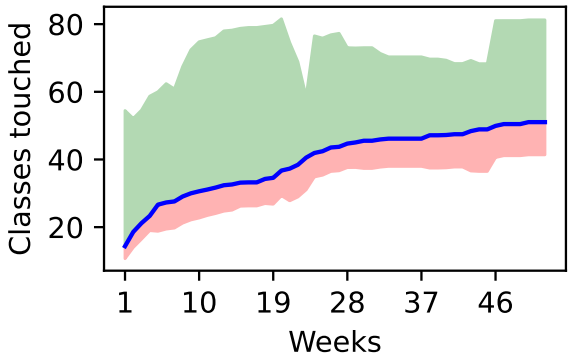
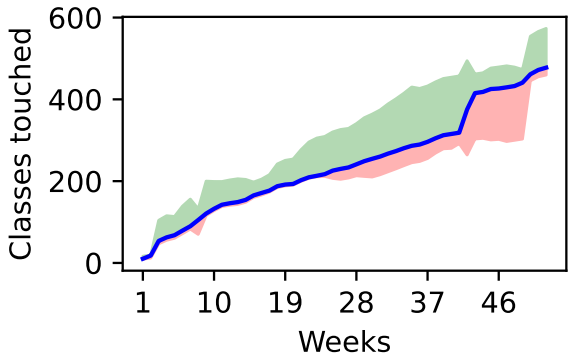


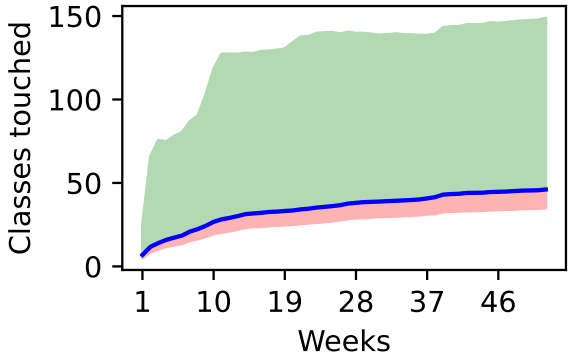
Figure 12: All developers (244) showing classes touched



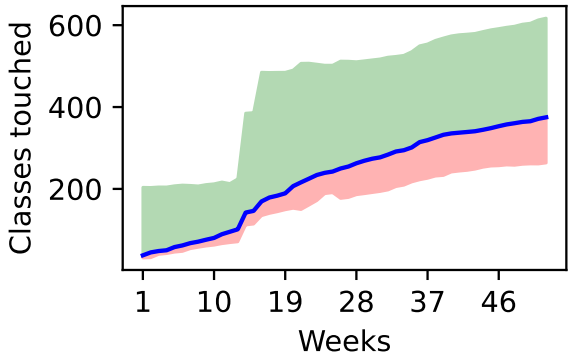
(a) Transient founder developers (13) showing classes touched



(b) Sustained founder developers (6) showing classes touched



(c) Transient later joiner developers (204) showing classes touched



(d) Sustained later joiner developers (21) showing classes touched

Figure 13: A time series of the number of week (x-axis) against the average (mean) total classes touched (y-axis), with positive (orange) and negative (red) filled standard deviation.

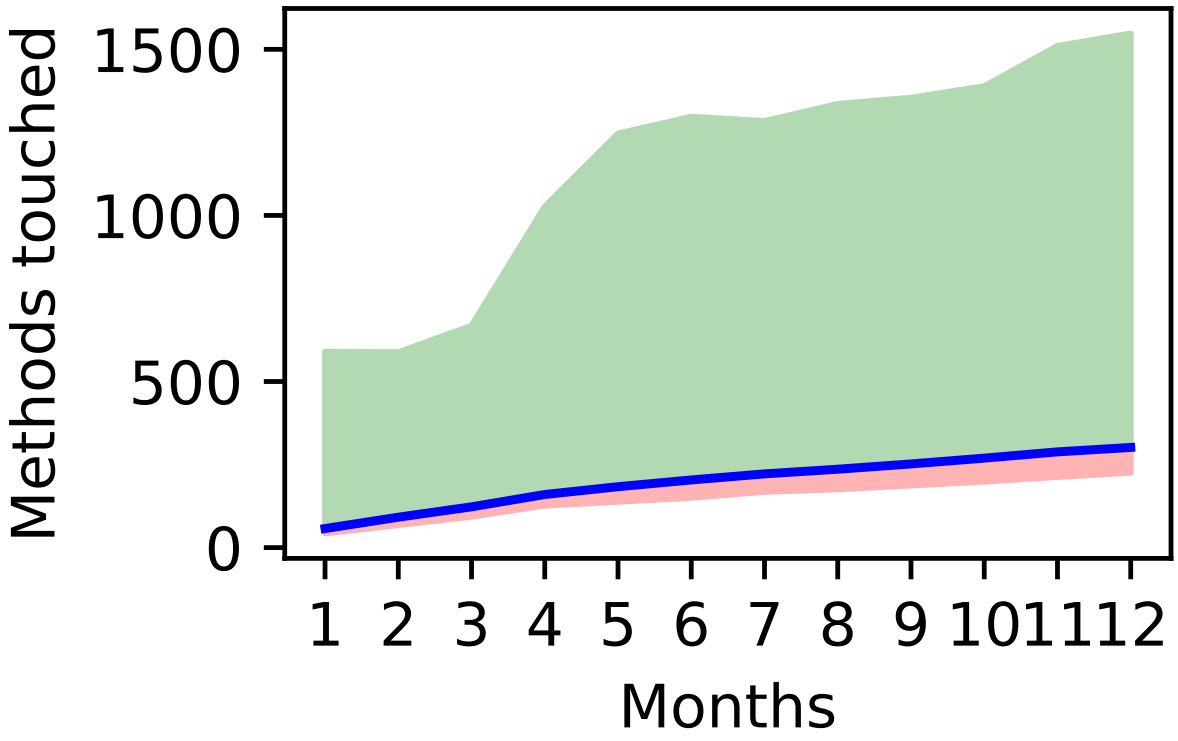
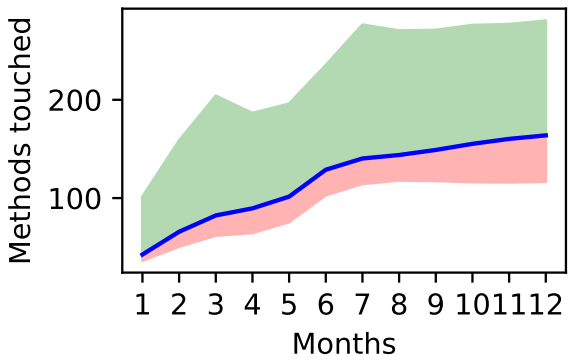
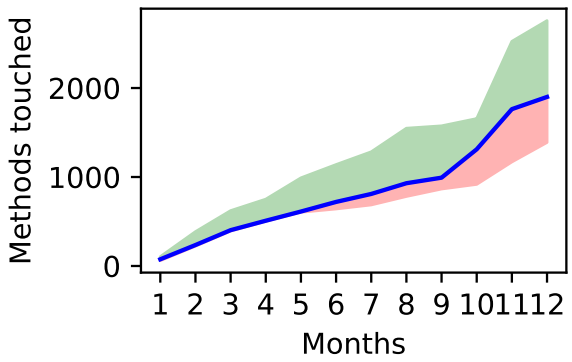


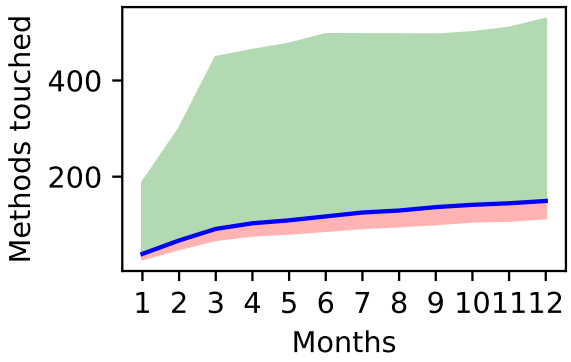
Figure 14: All developers (244) showing methods touched



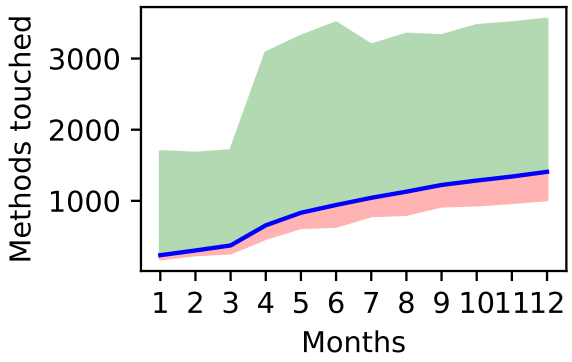
(a) Transient founder developers (13) showing methods touched



(b) Sustained founder developers (6) showing methods touched



(c) Transient later joiner developers (204) showing methods touched



(d) Sustained later joiner developers (21) showing methods touched

Figure 15: A time series of the number of month (x-axis) against the average (mean) total methods touched (y-axis), with positive (orange) and negative (red) filled standard deviation.

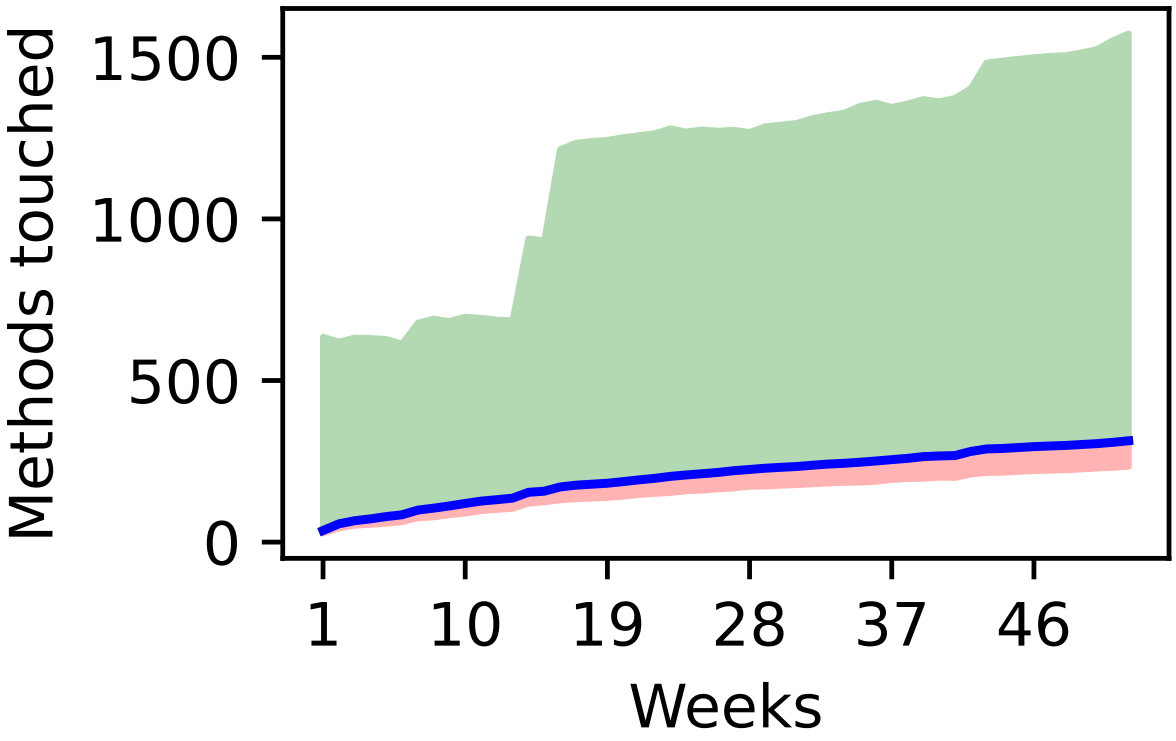
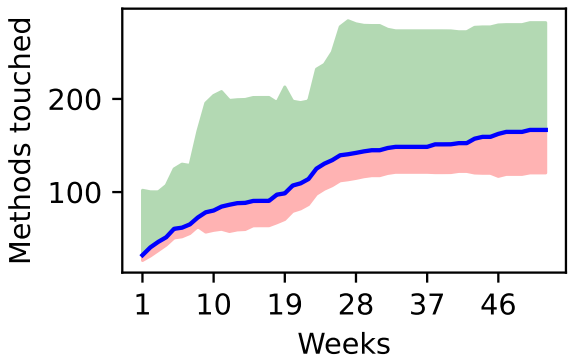
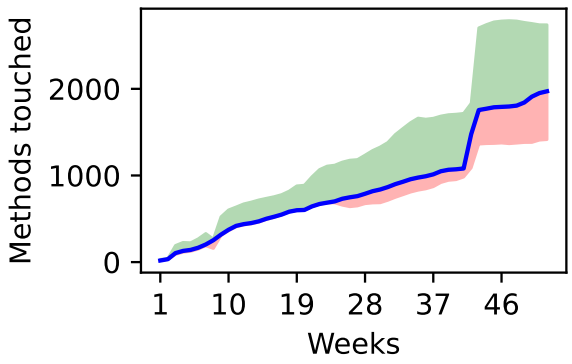


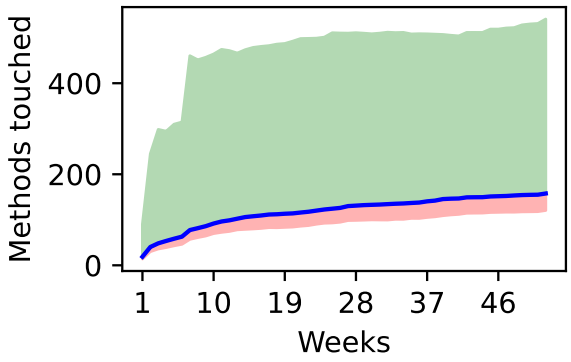
Figure 16: All developers (244) showing methods touched



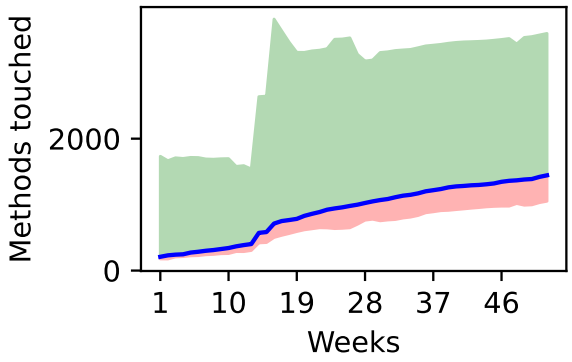
(a) Transient founder developers (13) showing methods touched



(b) Sustained founder developers (6) showing methods touched



(c) Transient later joiner developers (204) showing methods touched

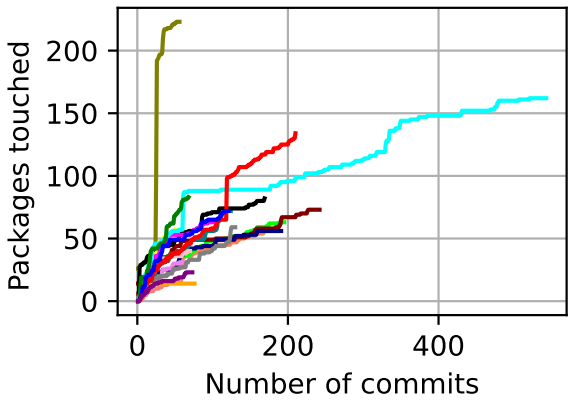


(d) Sustained later joiner developers (21) showing methods touched

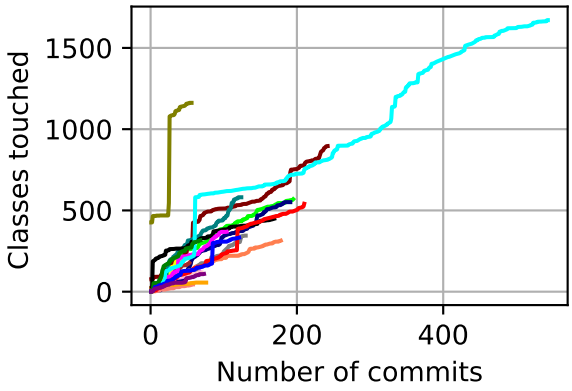
Figure 17: A time series of the number of week (x-axis) against the average (mean) total methods touched (y-axis), with positive (orange) and negative (red) filled standard deviation.

4.3 Scatter developer - 14 For developer commits and components touched.

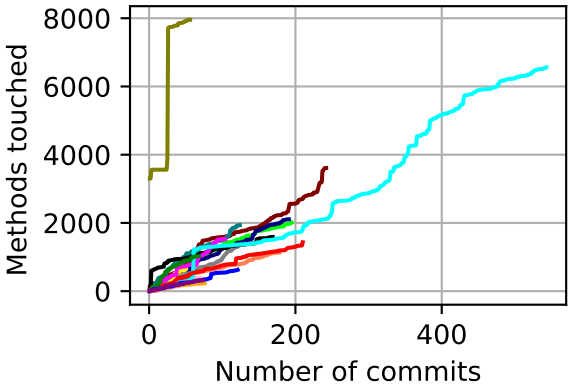
A scatter plot for the first 20 sustained late joiner developers showing the number of commits to components touched.



(a) packages



(b) classes



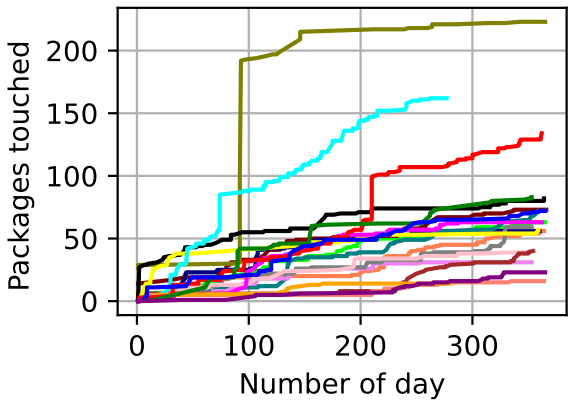
(c) methods

Figure 18: First twenty sustained later joiner developers from this repository. The number of commits against the number of components touched.

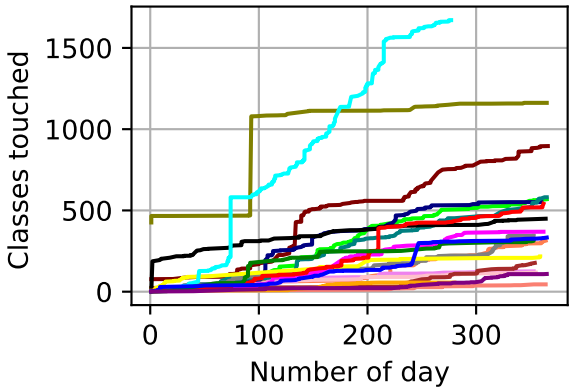
- Developer: 1774 in colour Coral. Developer: 2986 in colour Lime.
- Developer: 3000 in colour Maroon. Developer: 3020 in colour Navy.
- Developer: 3037 in colour Teal. Developer: 3050 in colour Olive.
- Developer: 3071 in colour Violet. Developer: 3072 in colour Gray.
- Developer: 3141 in colour Pink. Developer: 3143 in colour Brown.
- Developer: 3163 in colour Black. Developer: 3172 in colour Yellow.
- Developer: 3178 in colour Magenta. Developer: 3216 in colour Cyan.
- Developer: 3217 in colour Salmon. Developer: 3272 in colour Orange.
- Developer: 3339 in colour Red. Developer: 3354 in colour Green.
- Developer: 3456 in colour Blue. Developer: 3473 in colour Purple.

4.4 Scatter developer - 14 For developer day and components touched.

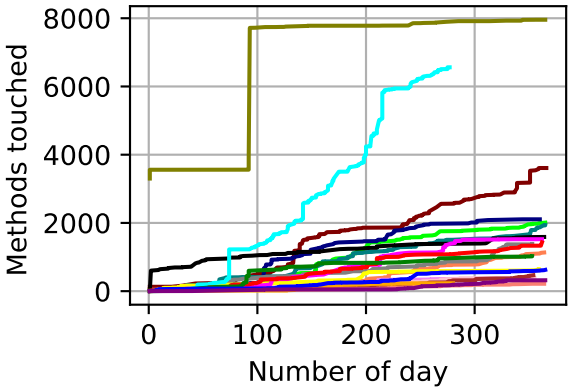
A scatter plot for the first 20 sustained late joiner developers showing the number of day to components touched.



(a) packages



(b) classes

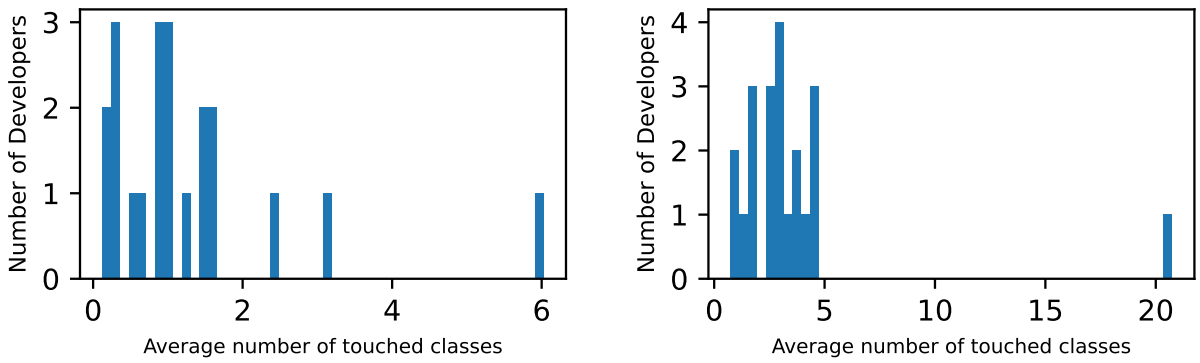


(c) methods

Figure 19: First twenty sustained later joiner developers from this repository. The number of day against the number of components touched.

- Developer: 1774 in colour Coral. Developer: 2986 in colour Lime.
- Developer: 3000 in colour Maroon. Developer: 3020 in colour Navy.
- Developer: 3037 in colour Teal. Developer: 3050 in colour Olive.
- Developer: 3071 in colour Violet. Developer: 3072 in colour Gray.
- Developer: 3141 in colour Pink. Developer: 3143 in colour Brown.
- Developer: 3163 in colour Black. Developer: 3172 in colour Yellow.
- Developer: 3178 in colour Magenta. Developer: 3216 in colour Cyan.
- Developer: 3217 in colour Salmon. Developer: 3272 in colour Orange.
- Developer: 3339 in colour Red. Developer: 3354 in colour Green.
- Developer: 3456 in colour Blue. Developer: 3473 in colour Purple.

4.5 Repository histogram commit - 14 For classes with total

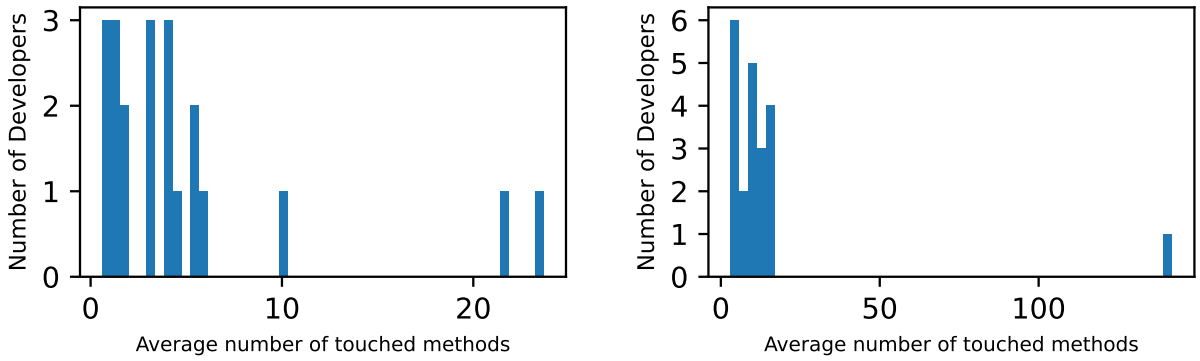


(a) By Day.

(b) By Commit.

Figure 20: Histogram of new classes touched (x-axis) against number of developers (y-axis). Graphs for 21 sustained later joiner developers joining the project after six months and contributed at least 50 commits over a period of 250 days or more.

4.6 Repository histogram commit - 14 For methods with total



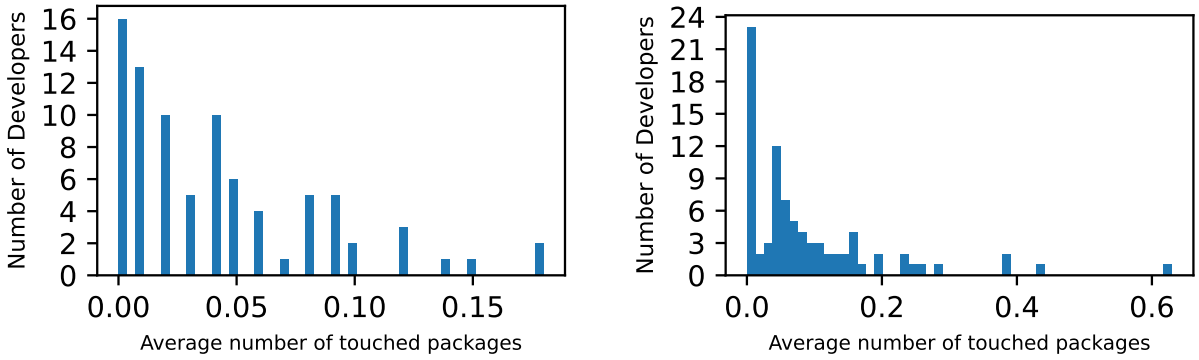
(a) By Day.

(b) By Commit.

Figure 21: Histogram of new methods touched (x-axis) against number of developers (y-axis). Graphs for 21 sustained later joiner developers joining the project after six months and contributed at least 50 commits over a period of 250 days or more.

5 Repository: 21

5.1 Repository histogram commit - 21 For packages with total



(a) By Day.

(b) By Commit.

Figure 22: Histogram of new packages touched (x-axis) against number of developers (y-axis). Graphs for 84 sustained later joiner developers joining the project after six months and contributed at least 50 commits over a period of 250 days or more.

5.2 Time series developer - 21 For packages touched for each period

A time series of packages touched on average each month.

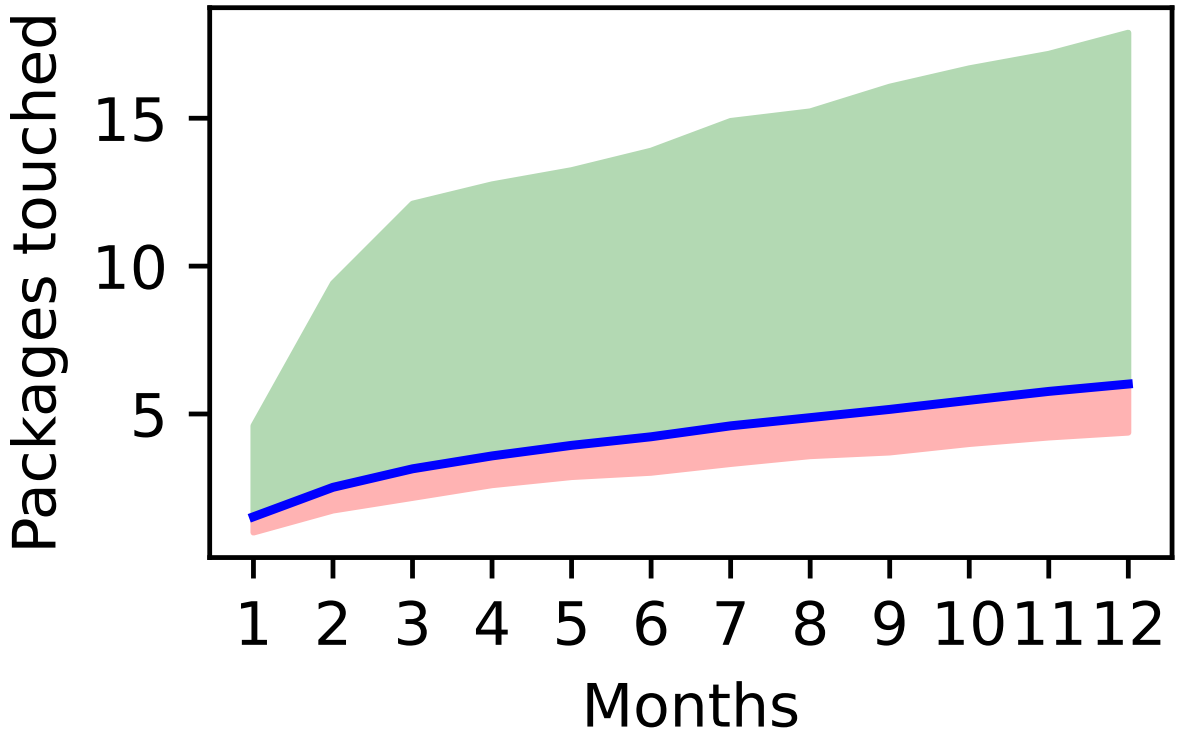
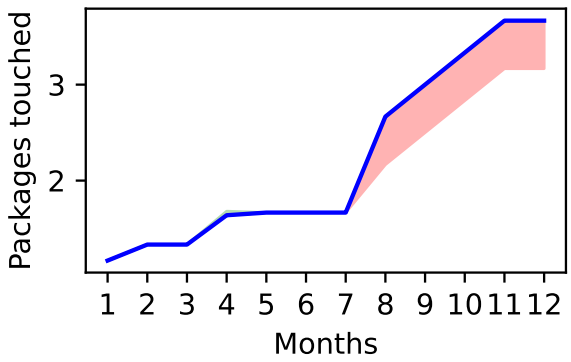
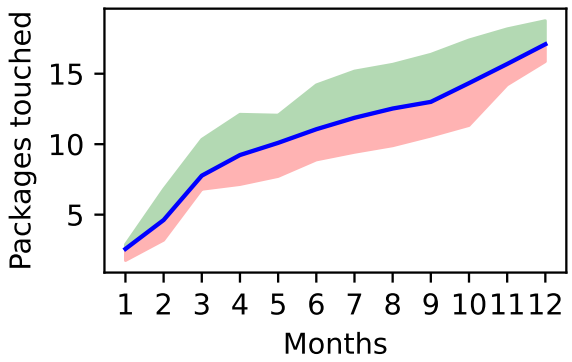


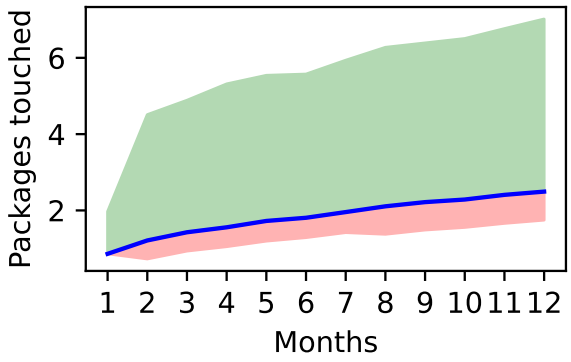
Figure 23: All developers (300) showing packages touched



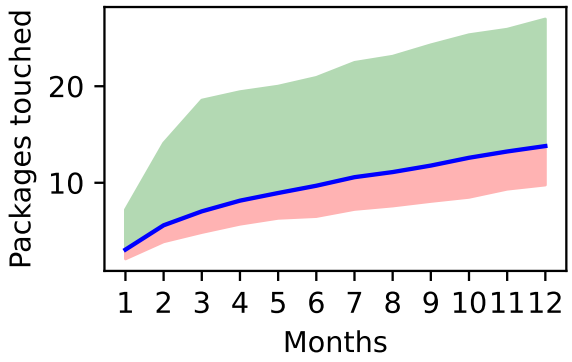
(a) Transient founder developers (3) showing packages touched



(b) Sustained founder developers (7) showing packages touched



(c) Transient later joiner developers (206) showing packages touched



(d) Sustained later joiner developers (84) showing packages touched

Figure 24: A time series of the number of month (x-axis) against the average (mean) total packages touched (y-axis), with positive (orange) and negative (red) filled standard deviation.

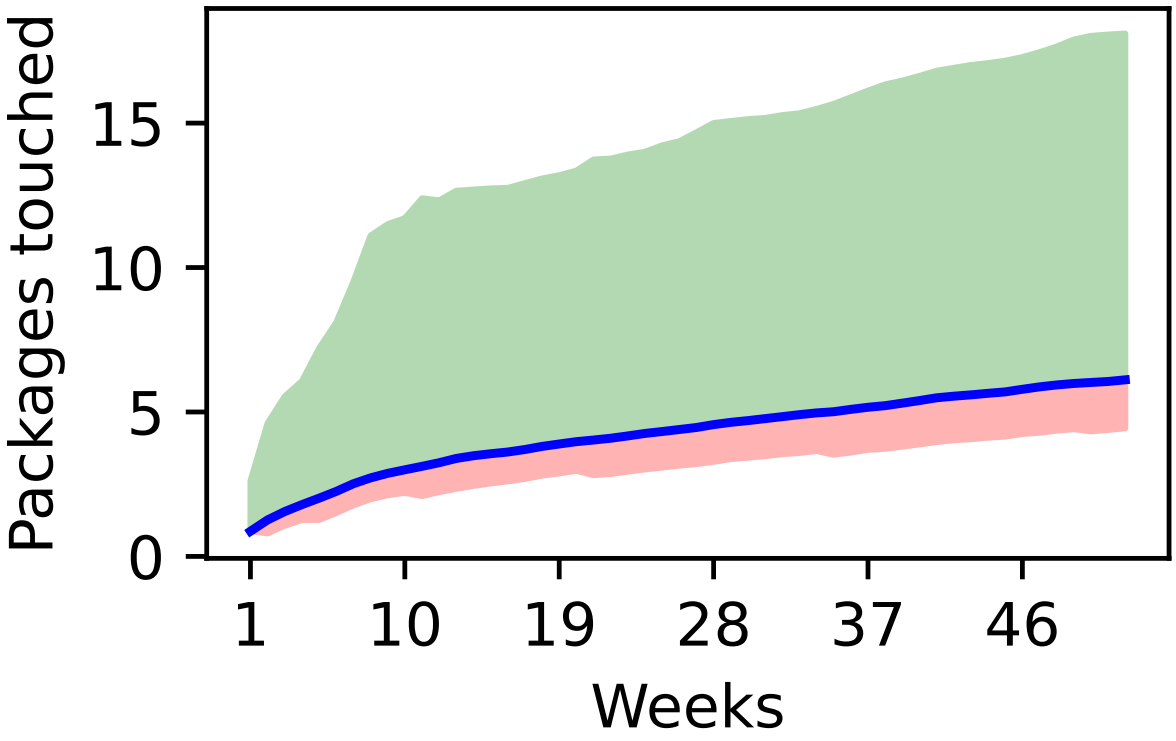
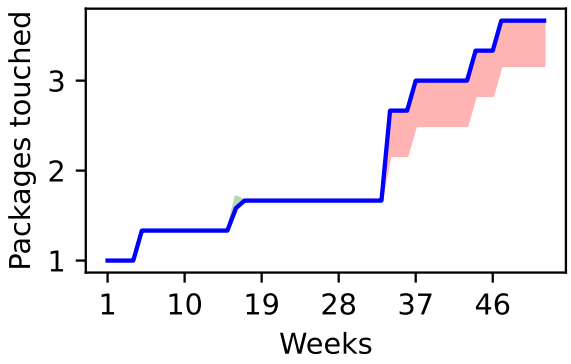
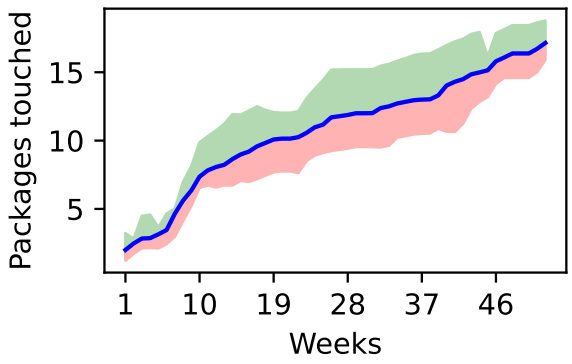


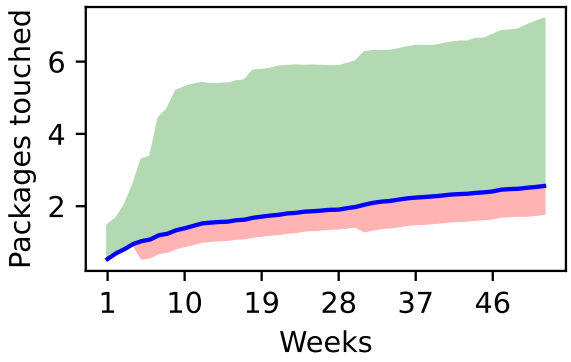
Figure 25: All developers (300) showing packages touched



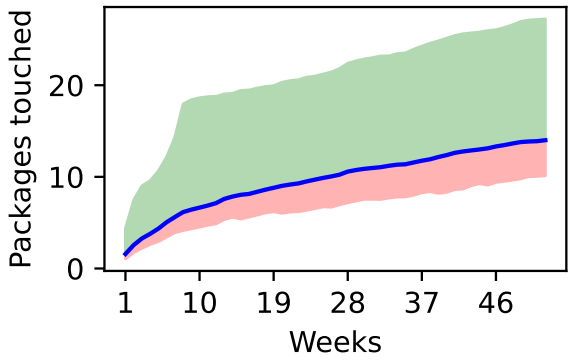
(a) Transient founder developers (3) showing packages touched



(b) Sustained founder developers (7) showing packages touched



(c) Transient later joiner developers (206) showing packages touched



(d) Sustained later joiner developers (84) showing packages touched

Figure 26: A time series of the number of week (x-axis) against the average (mean) total packages touched (y-axis), with positive (orange) and negative (red) filled standard deviation.

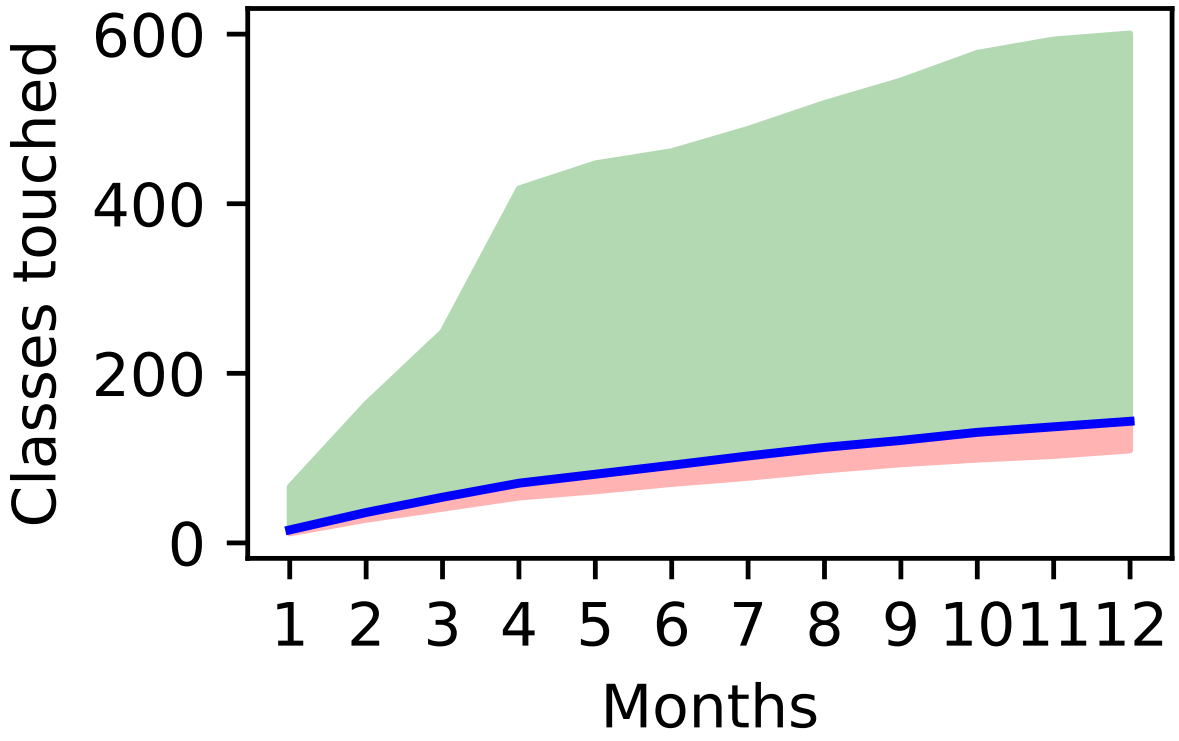
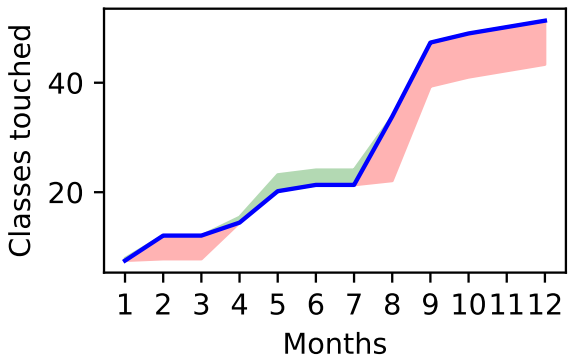
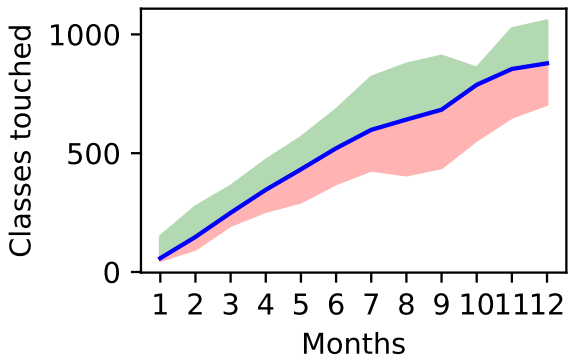


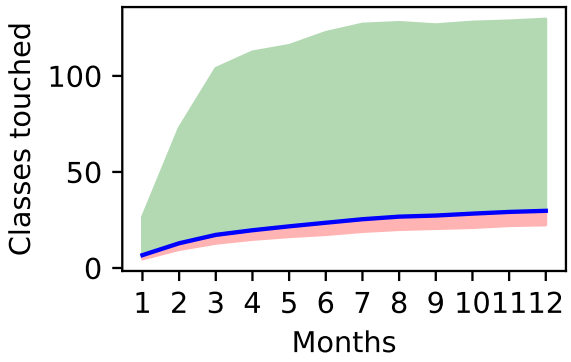
Figure 27: All developers (300) showing classes touched



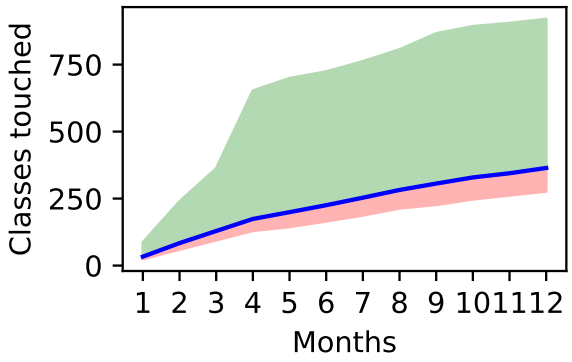
(a) Transient founder developers (3) showing classes touched



(b) Sustained founder developers (7) showing classes touched



(c) Transient later joiner developers (206) showing classes touched



(d) Sustained later joiner developers (84) showing classes touched

Figure 28: A time series of the number of month (x-axis) against the average (mean) total classes touched (y-axis), with positive (orange) and negative (red) filled standard deviation.

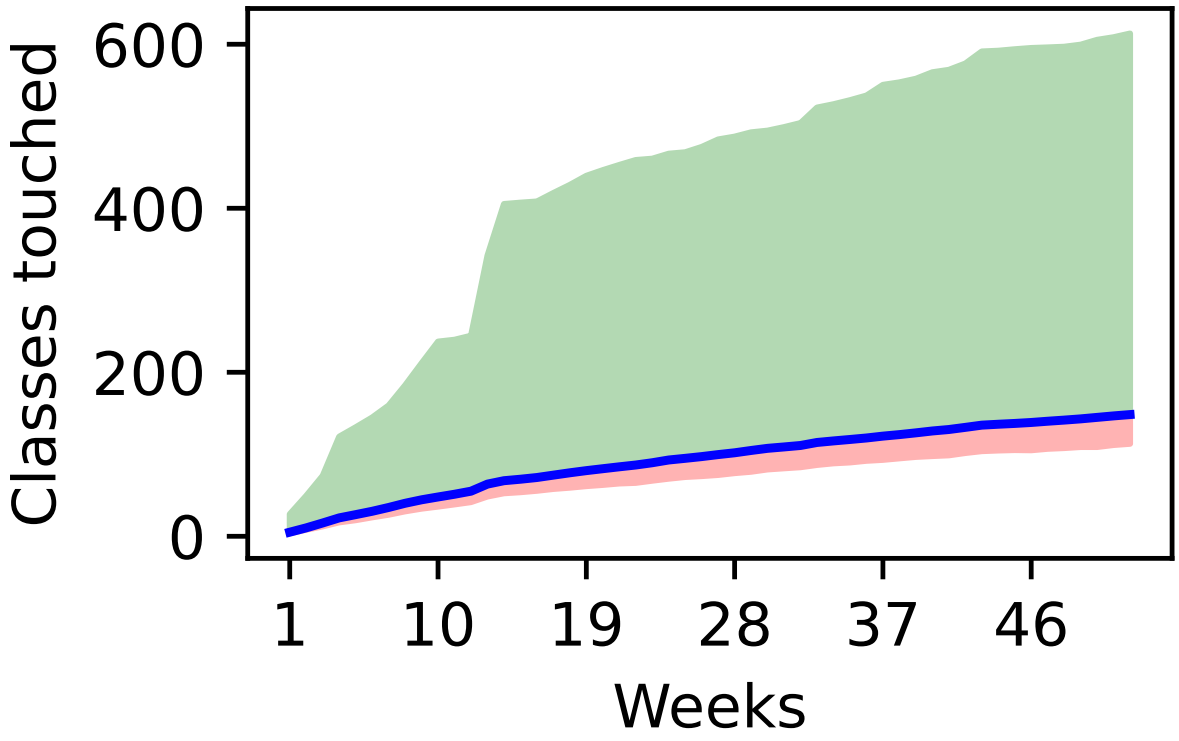
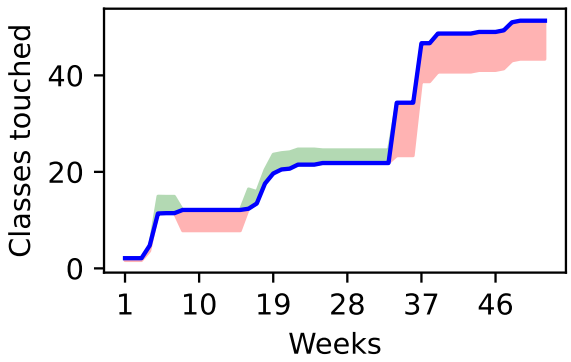
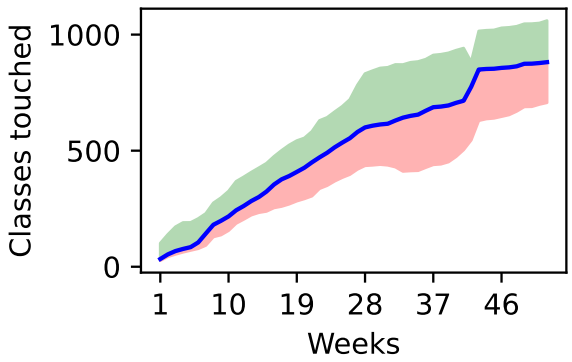


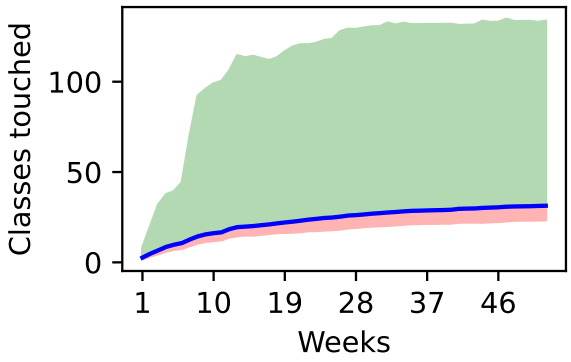
Figure 29: All developers (300) showing classes touched



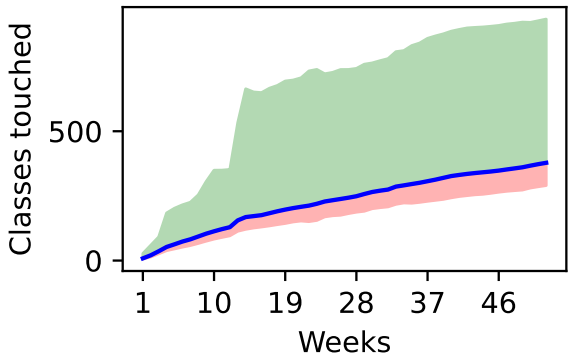
(a) Transient founder developers (3) showing classes touched



(b) Sustained founder developers (7) showing classes touched



(c) Transient later joiner developers (206) showing classes touched



(d) Sustained later joiner developers (84) showing classes touched

Figure 30: A time series of the number of week (x-axis) against the average (mean) total classes touched (y-axis), with positive (orange) and negative (red) filled standard deviation.

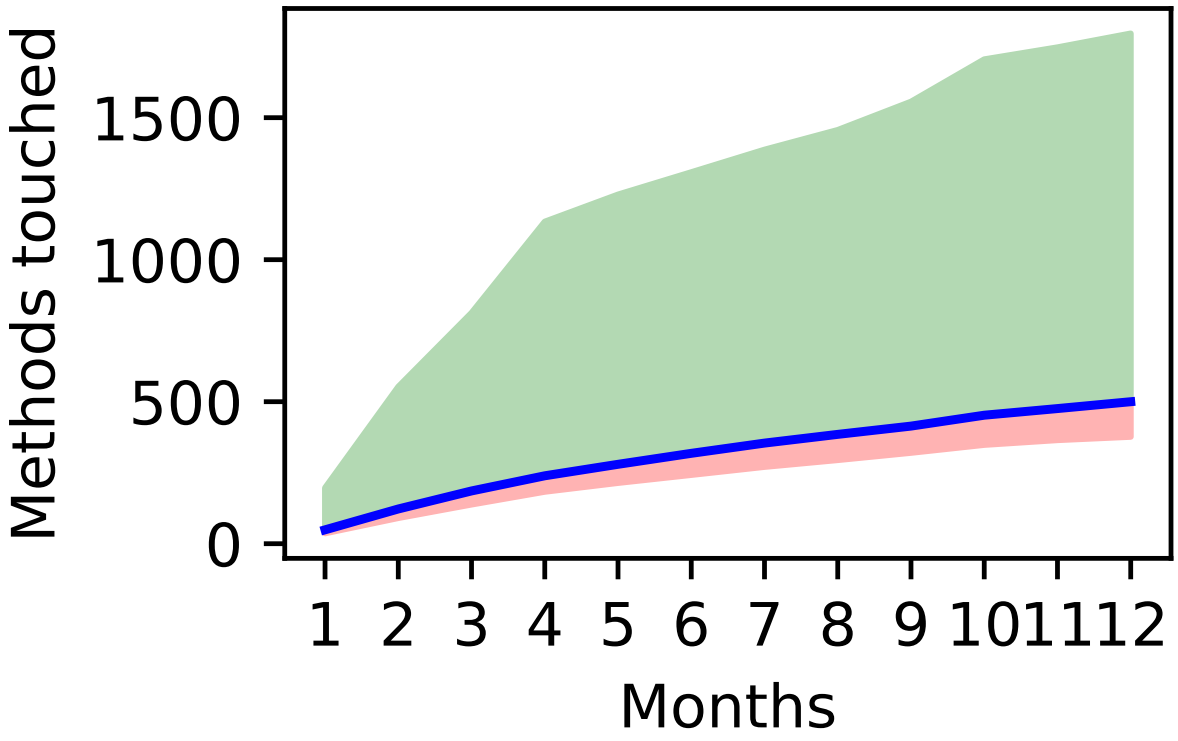
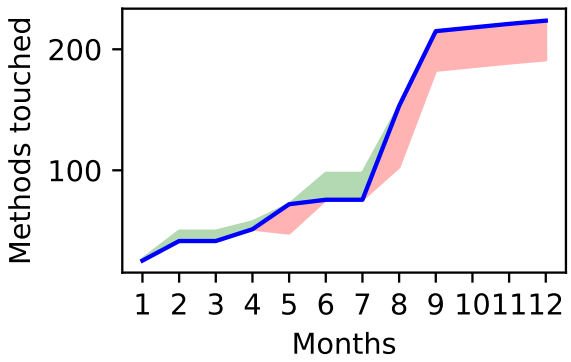
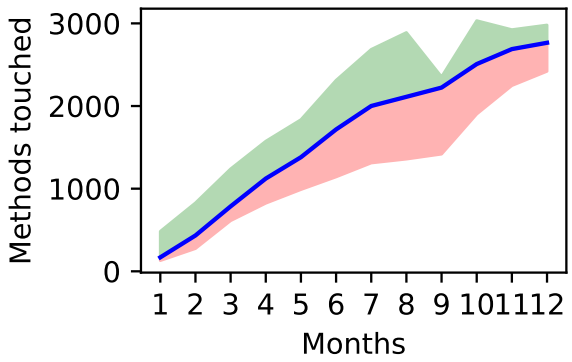


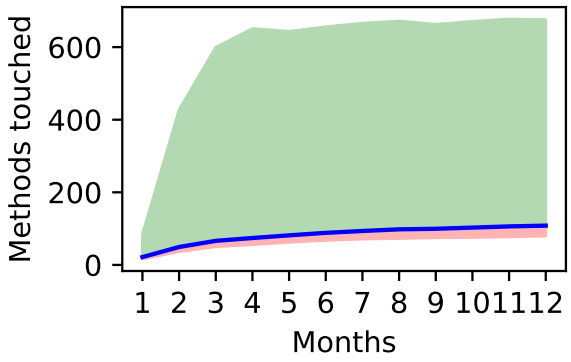
Figure 31: All developers (300) showing methods touched



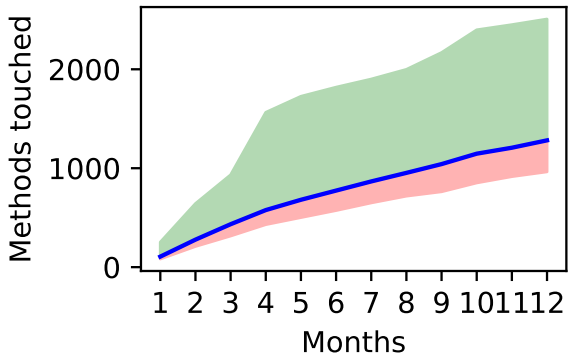
(a) Transient founder developers (3) showing methods touched



(b) Sustained founder developers (7) showing methods touched



(c) Transient later joiner developers (206) showing methods touched



(d) Sustained later joiner developers (84) showing methods touched

Figure 32: A time series of the number of month (x-axis) against the average (mean) total methods touched (y-axis), with positive (orange) and negative (red) filled standard deviation.

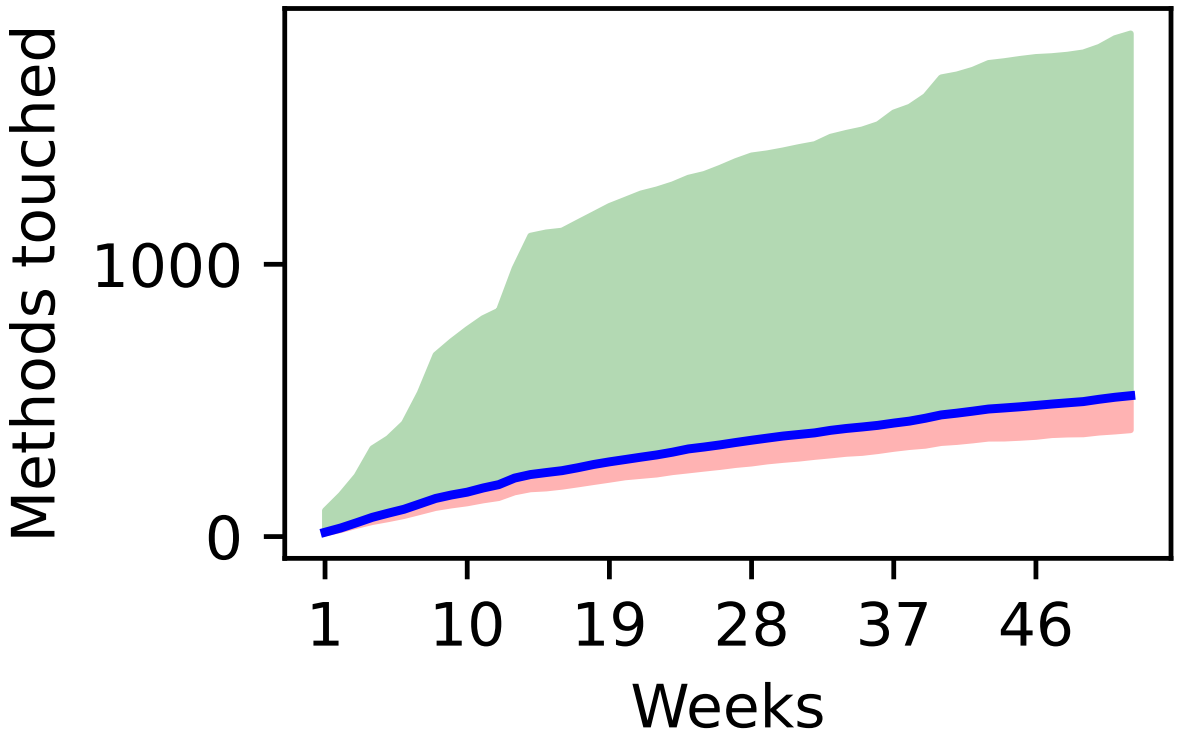
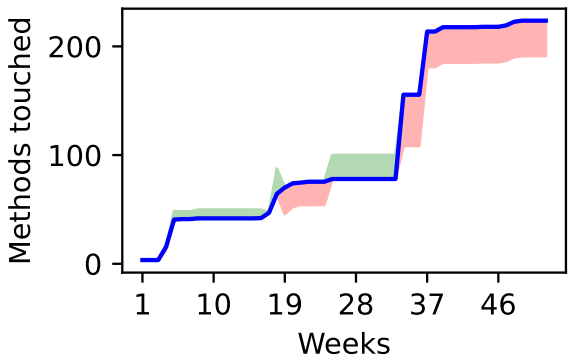
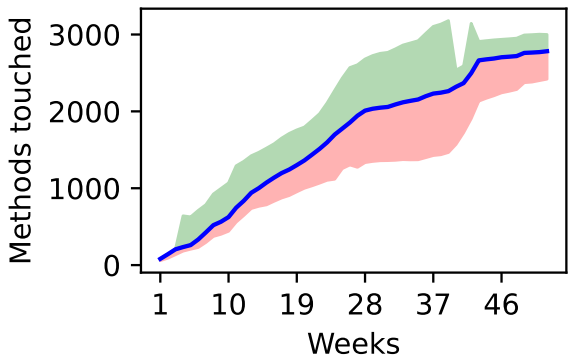


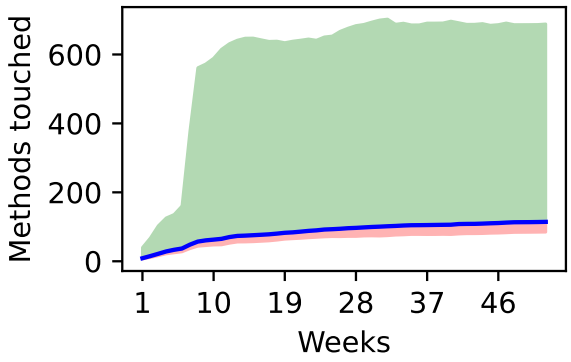
Figure 33: All developers (300) showing methods touched



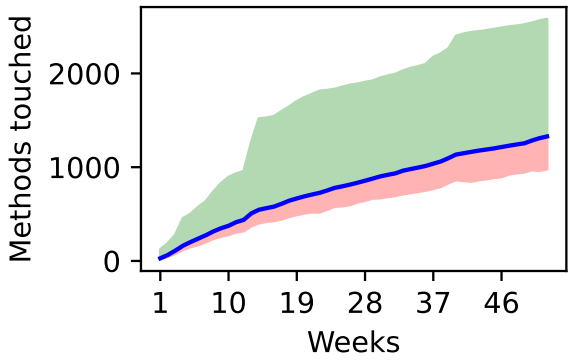
(a) Transient founder developers (3) showing methods touched



(b) Sustained founder developers (7) showing methods touched



(c) Transient later joiner developers (206) showing methods touched

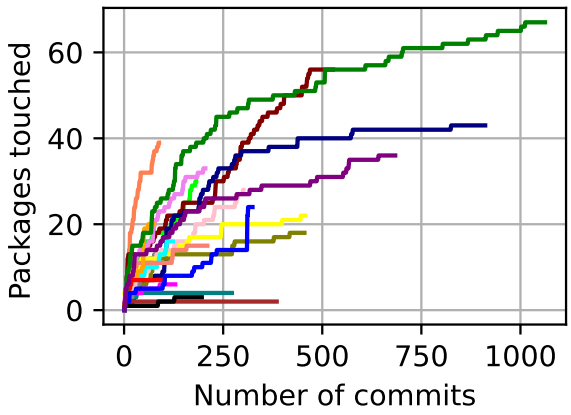


(d) Sustained later joiner developers (84) showing methods touched

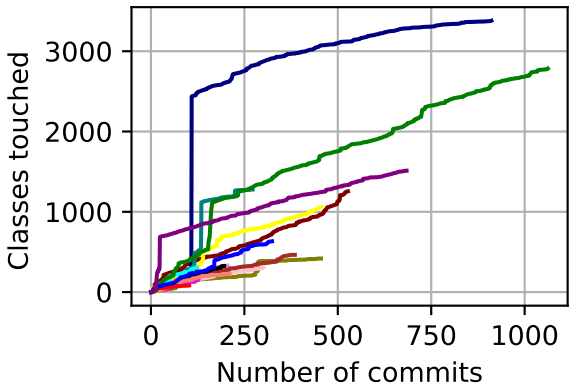
Figure 34: A time series of the number of week (x-axis) against the average (mean) total methods touched (y-axis), with positive (orange) and negative (red) filled standard deviation.

5.3 Scatter developer - 21 For developer commits and components touched.

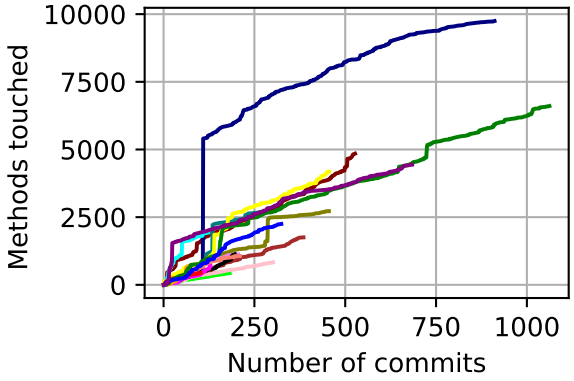
A scatter plot for the first 20 sustained late joiner developers showing the number of commits to components touched.



(a) packages



(b) classes



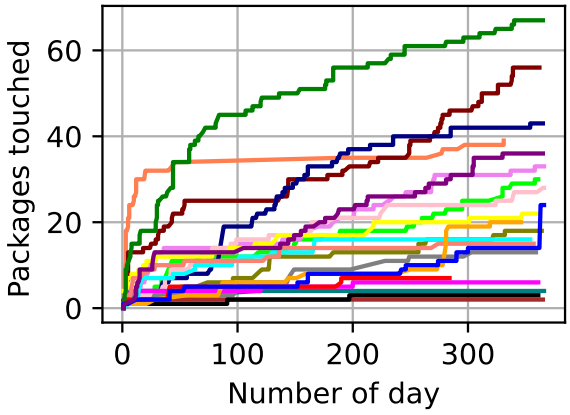
(c) methods

Figure 35: First twenty sustained later joiner developers from this repository. The number of commits against the number of components touched.

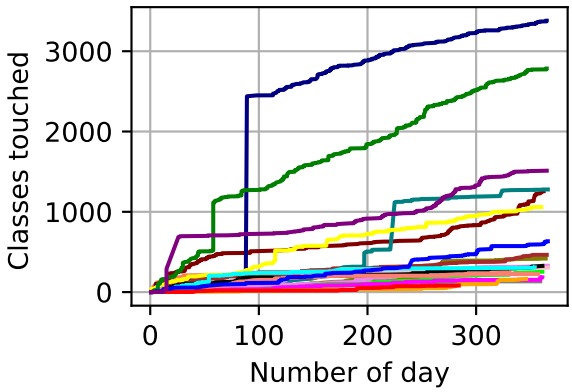
- Developer: 199 in colour Coral. Developer: 6252 in colour Lime.
- Developer: 6253 in colour Maroon. Developer: 6259 in colour Navy.
- Developer: 6260 in colour Teal. Developer: 6262 in colour Olive.
- Developer: 6263 in colour Violet. Developer: 6264 in colour Gray.
- Developer: 6267 in colour Pink. Developer: 6271 in colour Brown.
- Developer: 6272 in colour Black. Developer: 6276 in colour Yellow.
- Developer: 6278 in colour Magenta. Developer: 6280 in colour Cyan.
- Developer: 6282 in colour Salmon. Developer: 6284 in colour Orange.
- Developer: 6288 in colour Red. Developer: 6295 in colour Green.
- Developer: 6297 in colour Blue. Developer: 6298 in colour Purple.

5.4 Scatter developer - 21 For developer day and components touched.

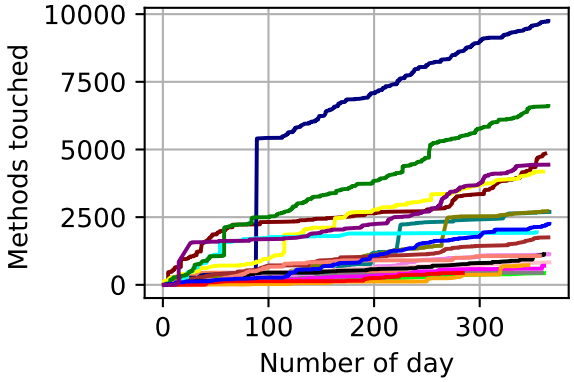
A scatter plot for the first 20 sustained late joiner developers showing the number of day to components touched.



(a) packages



(b) classes

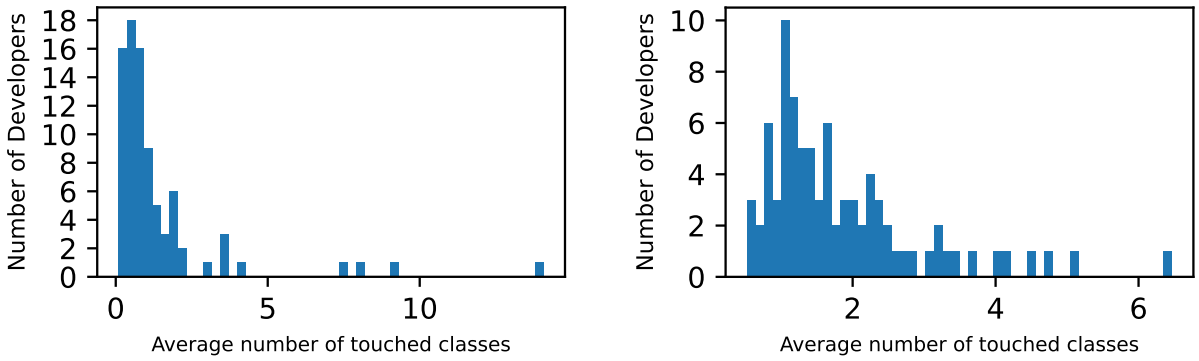


(c) methods

Figure 36: First twenty sustained later joiner developers from this repository. The number of day against the number of components touched.

- Developer: 199 in colour Coral. Developer: 6252 in colour Lime.
- Developer: 6253 in colour Maroon. Developer: 6259 in colour Navy.
- Developer: 6260 in colour Teal. Developer: 6262 in colour Olive.
- Developer: 6263 in colour Violet. Developer: 6264 in colour Gray.
- Developer: 6267 in colour Pink. Developer: 6271 in colour Brown.
- Developer: 6272 in colour Black. Developer: 6276 in colour Yellow.
- Developer: 6278 in colour Magenta. Developer: 6280 in colour Cyan.
- Developer: 6282 in colour Salmon. Developer: 6284 in colour Orange.
- Developer: 6288 in colour Red. Developer: 6295 in colour Green.
- Developer: 6297 in colour Blue. Developer: 6298 in colour Purple.

5.5 Repository histogram commit - 21 For classes with total

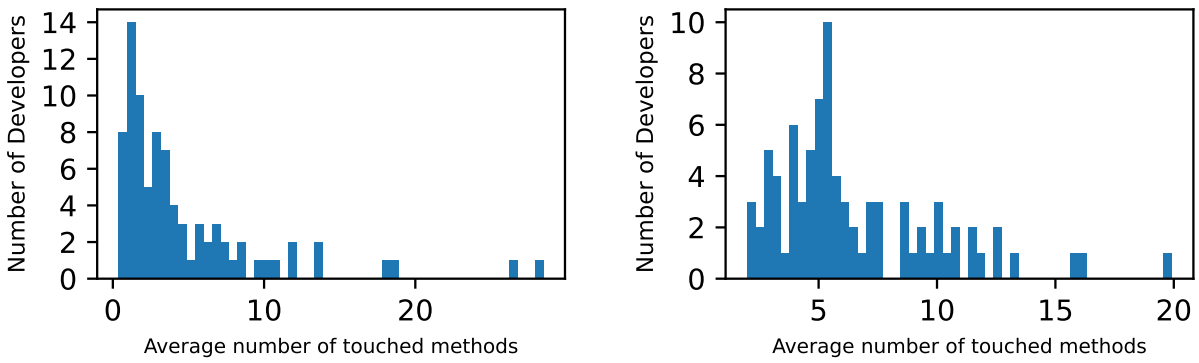


(a) By Day.

(b) By Commit.

Figure 37: Histogram of new classes touched (x-axis) against number of developers (y-axis). Graphs for 84 sustained later joiner developers joining the project after six months and contributed at least 50 commits over a period of 250 days or more.

5.6 Repository histogram commit - 21 For methods with total



(a) By Day.

(b) By Commit.

Figure 38: Histogram of new methods touched (x-axis) against number of developers (y-axis). Graphs for 84 sustained later joiner developers joining the project after six months and contributed at least 50 commits over a period of 250 days or more.

5.7 Box plot developer - 14 For packages touched for each period

A box plot of packages touched on average each period the number of commits to components touched.

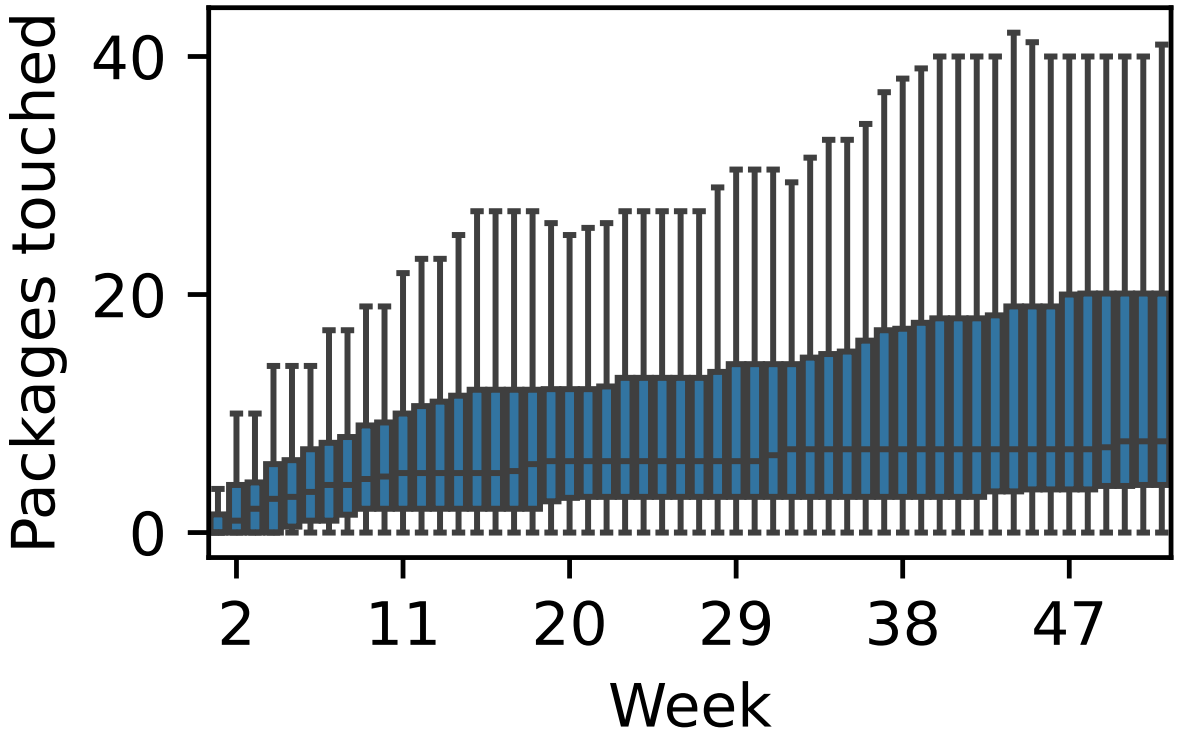
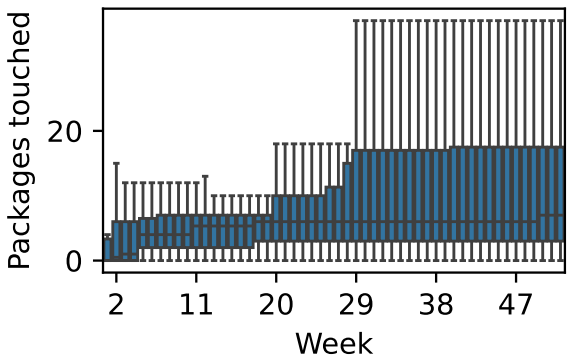
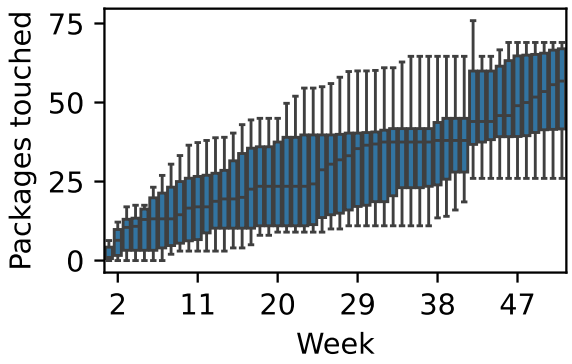


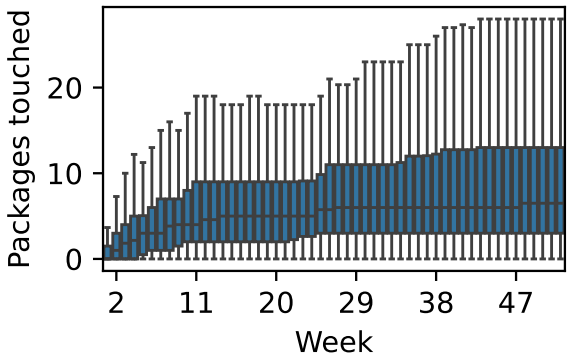
Figure 39: All developers (244) showing packages touched



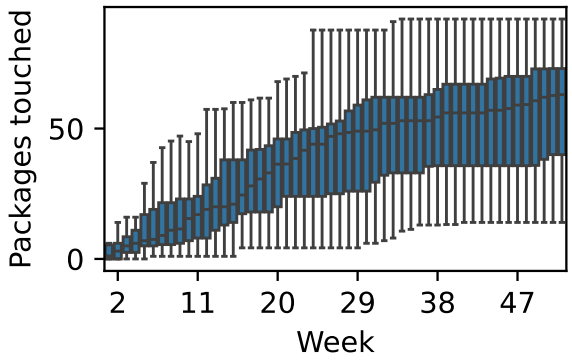
(a) Transient founder developers (13) showing packages touched



(b) Sustained founder developers (6) showing packages touched



(c) Transient later joiner developers (204) showing packages touched



(d) Sustained later joiner developers (21) showing packages touched

Figure 40: A box plot of total packages touched on mean each month, with quartile shading.

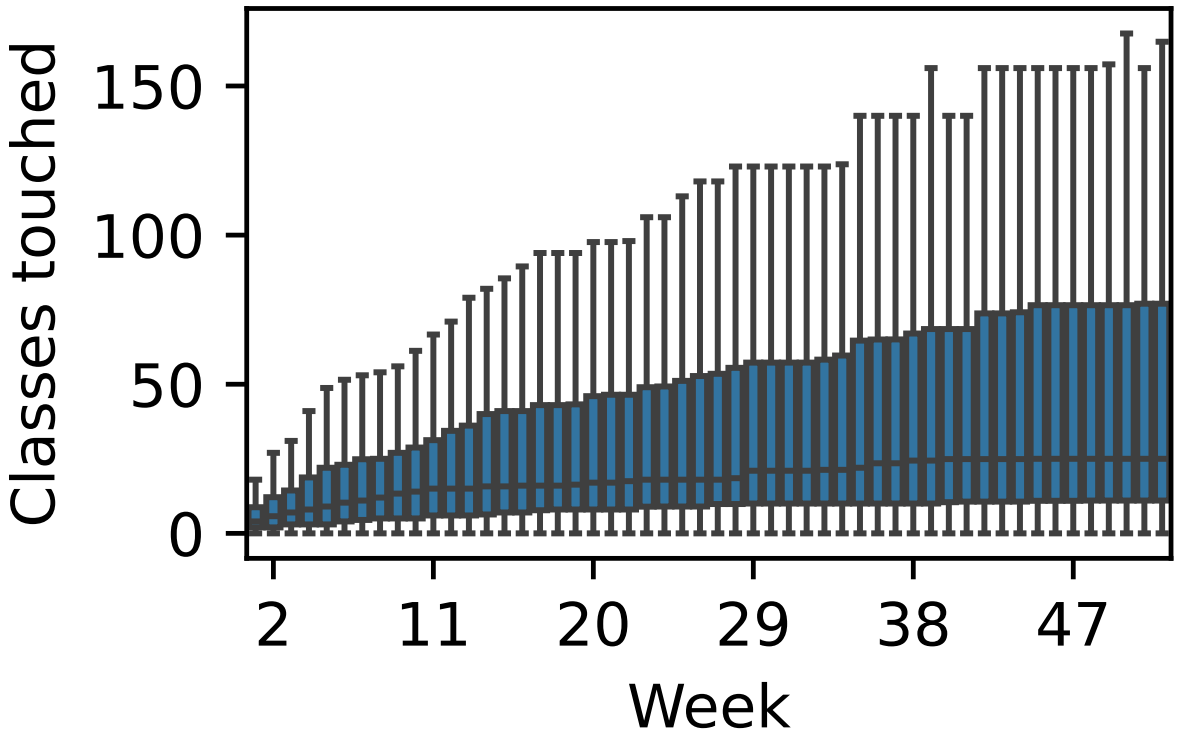
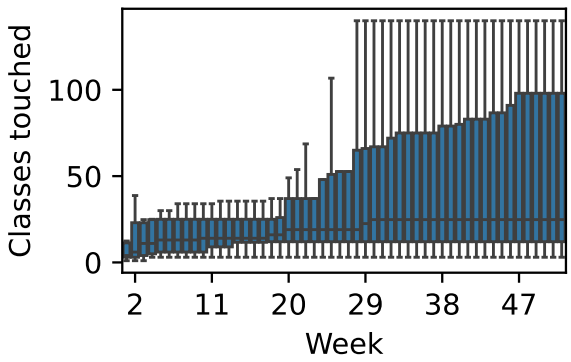
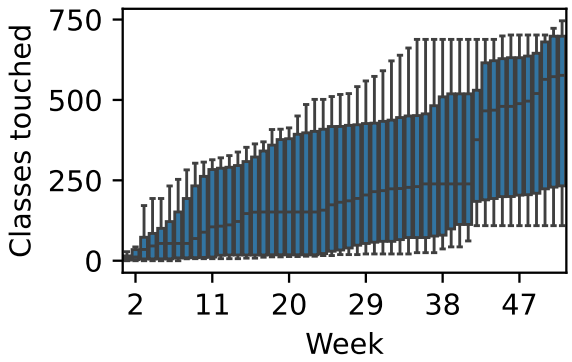


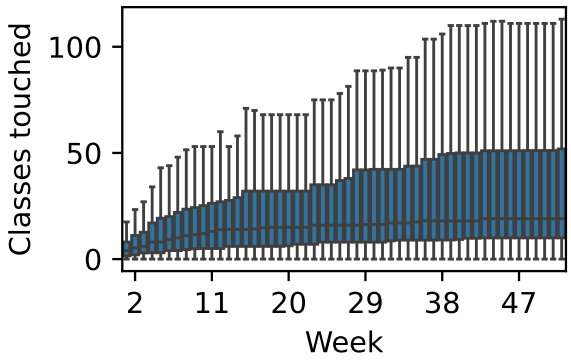
Figure 41: All developers (244) showing classes touched



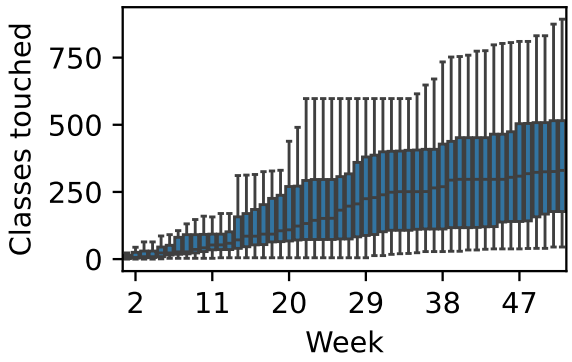
(a) Transient founder developers (13) showing classes touched



(b) Sustained founder developers (6) showing classes touched



(c) Transient later joiner developers (204) showing classes touched



(d) Sustained later joiner developers (21) showing classes touched

Figure 42: A box plot of total classes touched on mean each month, with quartile shading.

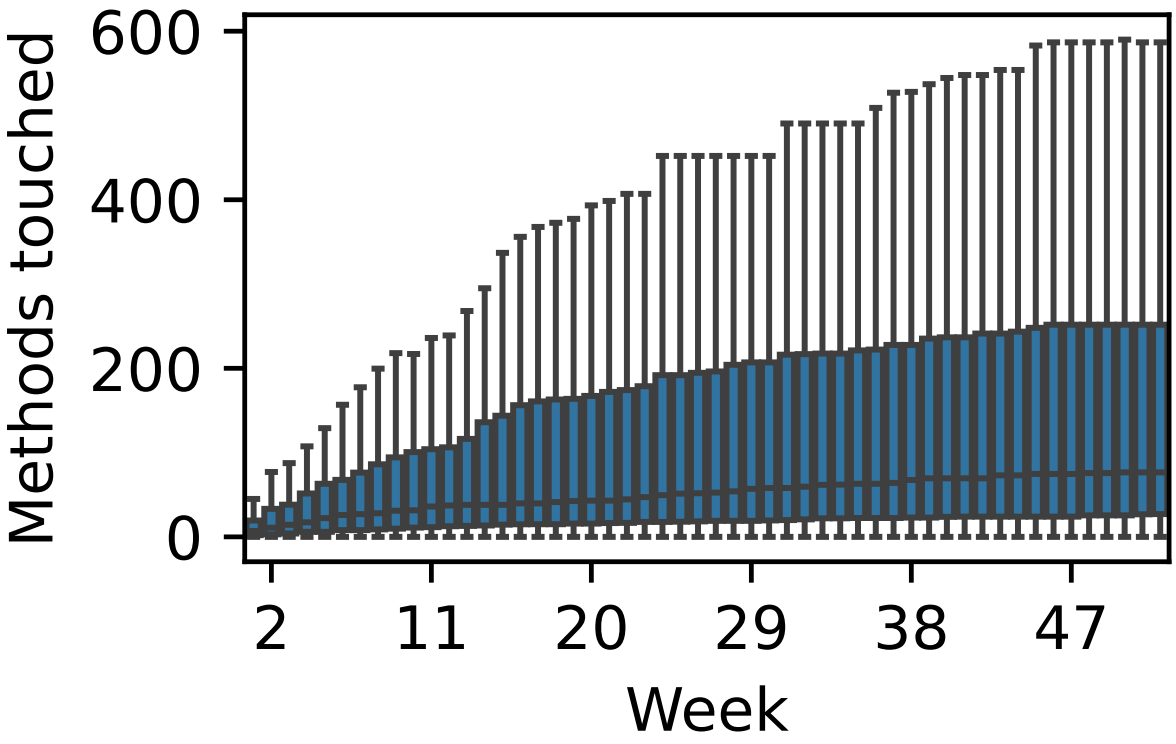
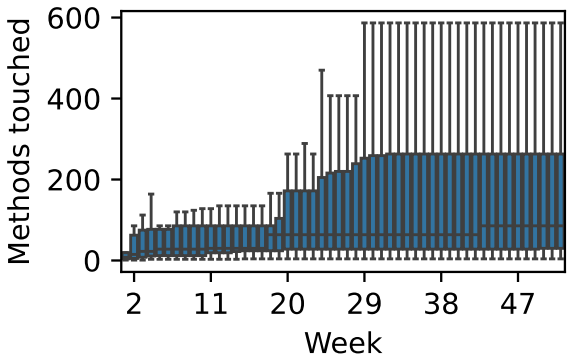
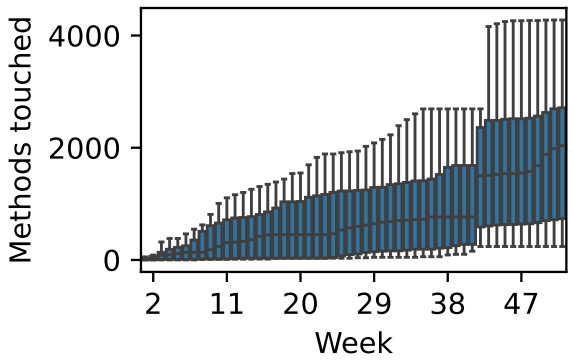


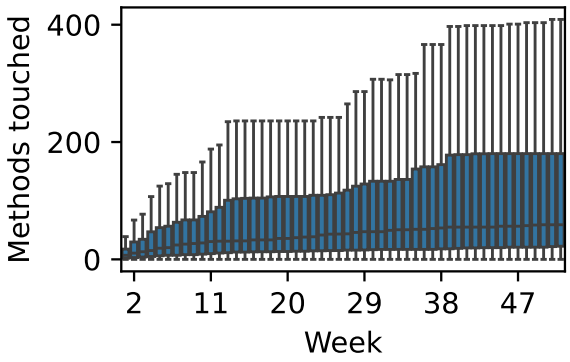
Figure 43: All developers (244) showing methods touched



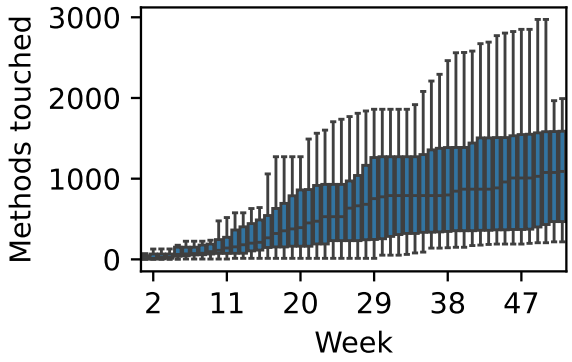
(a) Transient founder developers (13) showing methods touched



(b) Sustained founder developers (6) showing methods touched



(c) Transient later joiner developers (204) showing methods touched



(d) Sustained later joiner developers (21) showing methods touched

Figure 44: A box plot of total methods touched on mean each month, with quartile shading.

5.8 Box plot developer - 21 For packages touched for each period

A box plot of packages touched on average each period the number of commits to components touched.

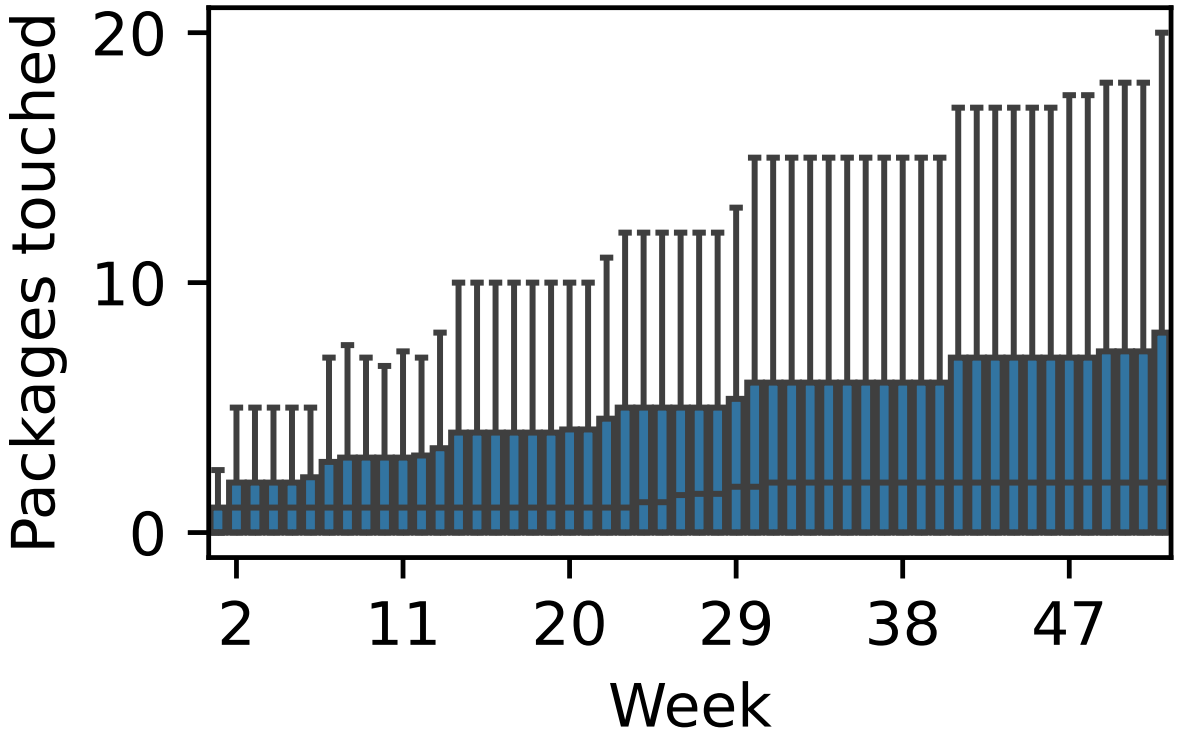
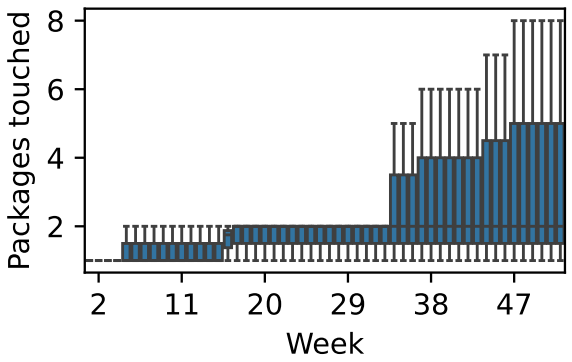
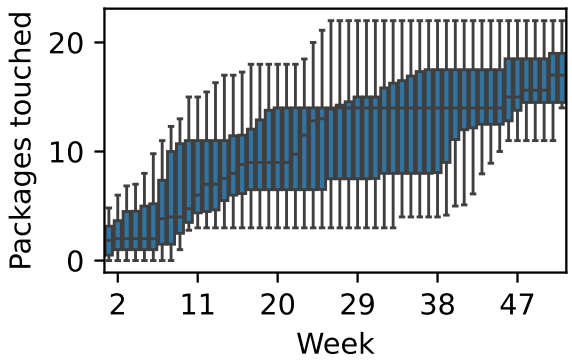


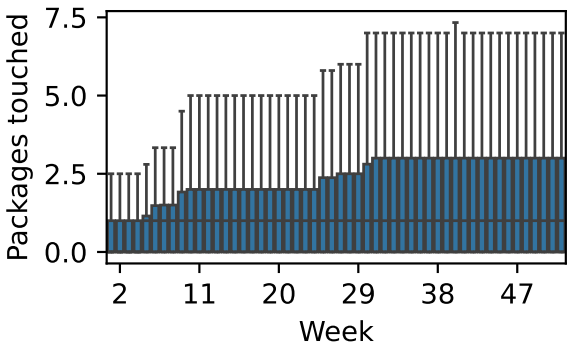
Figure 45: All developers (300) showing packages touched



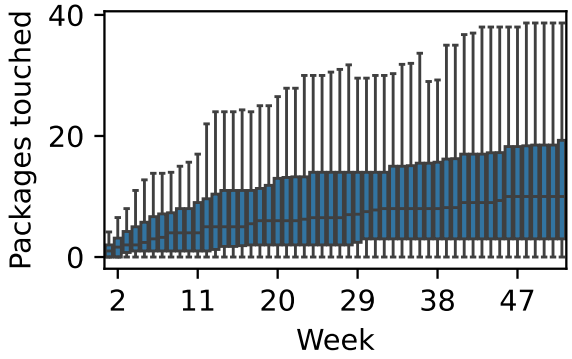
(a) Transient founder developers (3) showing packages touched



(b) Sustained founder developers (7) showing packages touched



(c) Transient later joiner developers (206) showing packages touched



(d) Sustained later joiner developers (84) showing packages touched

Figure 46: A box plot of total packages touched on mean each month, with quartile shading.

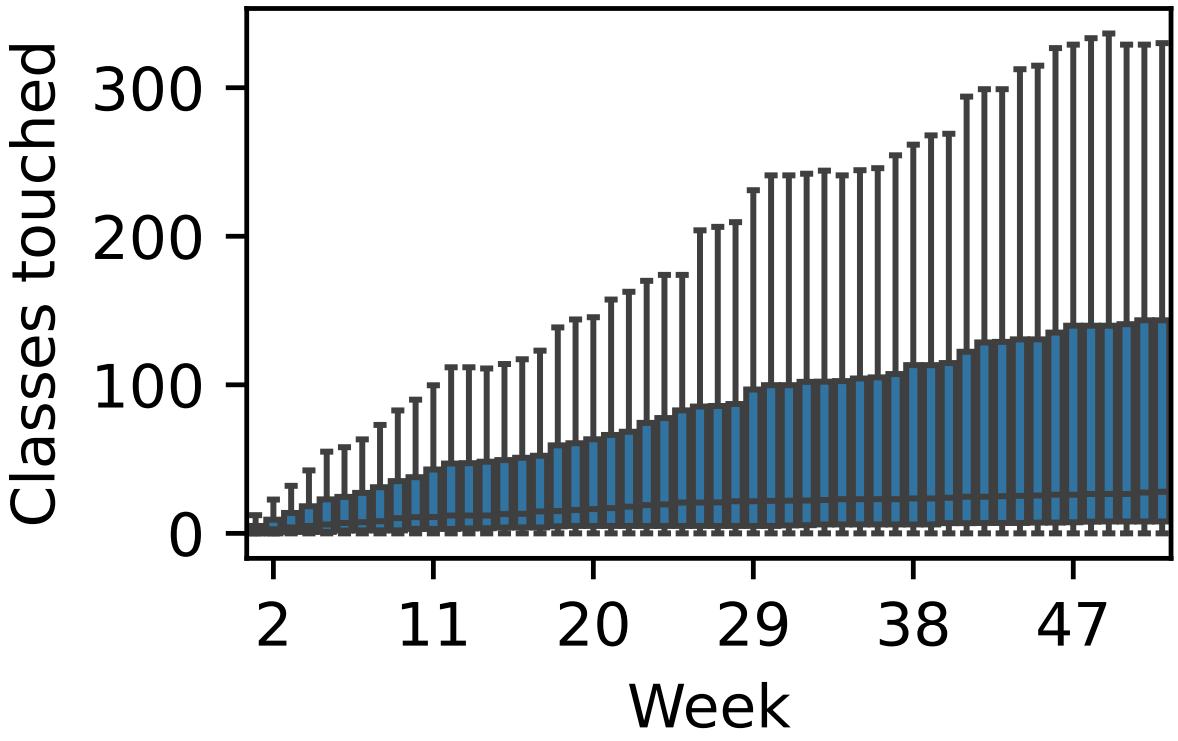
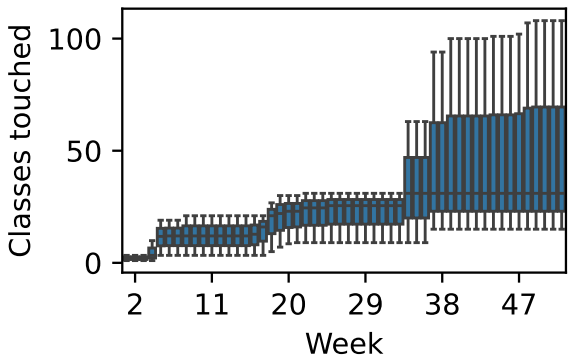
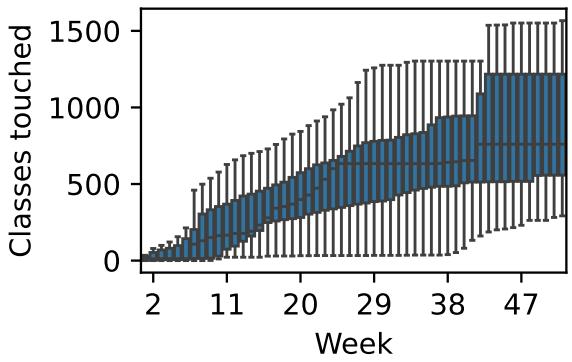


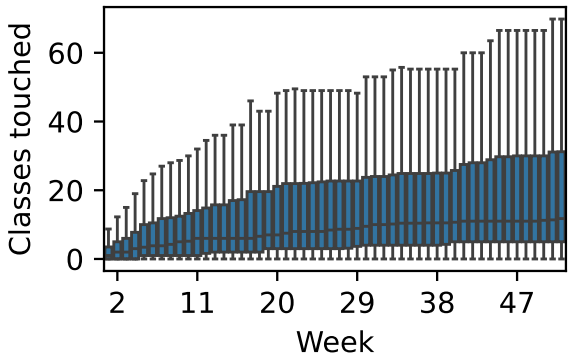
Figure 47: All developers (300) showing classes touched



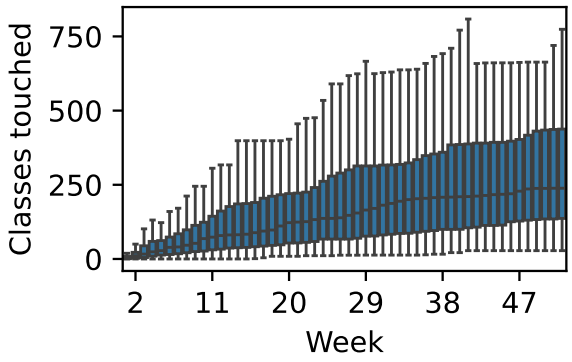
(a) Transient founder developers (3) showing classes touched



(b) Sustained founder developers (7) showing classes touched



(c) Transient later joiner developers (206) showing classes touched



(d) Sustained later joiner developers (84) showing classes touched

Figure 48: A box plot of total classes touched on mean each month, with quartile shading.

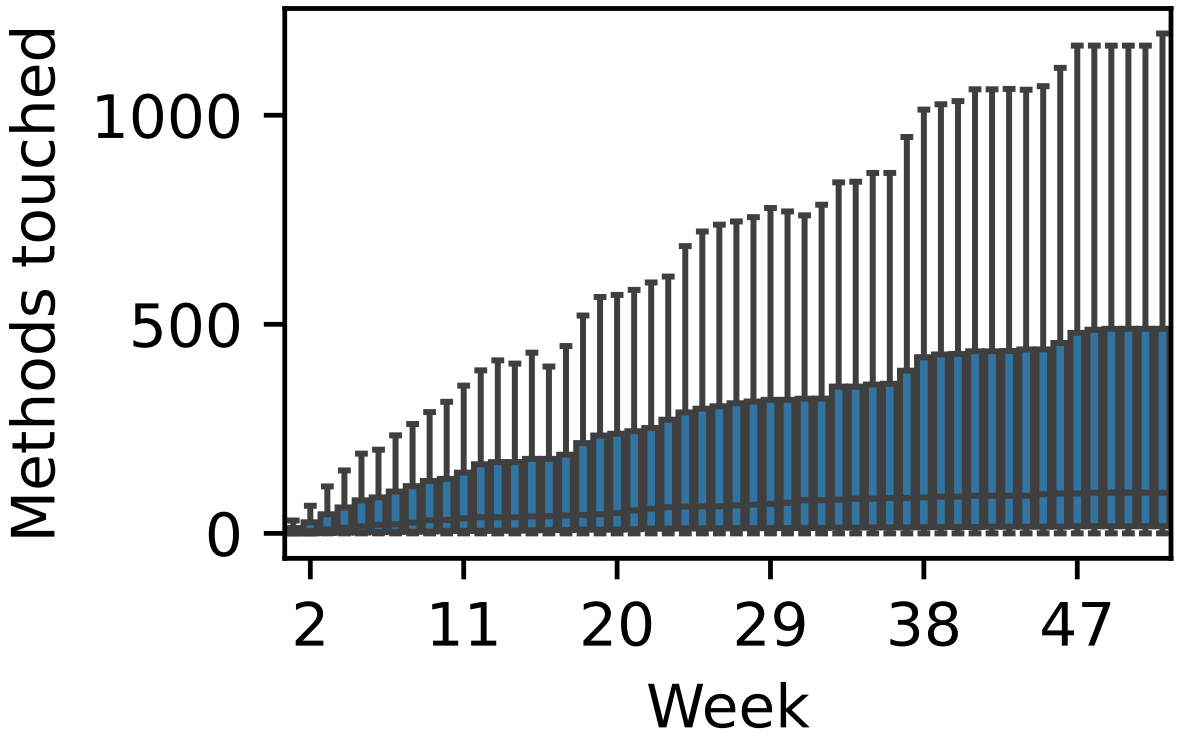
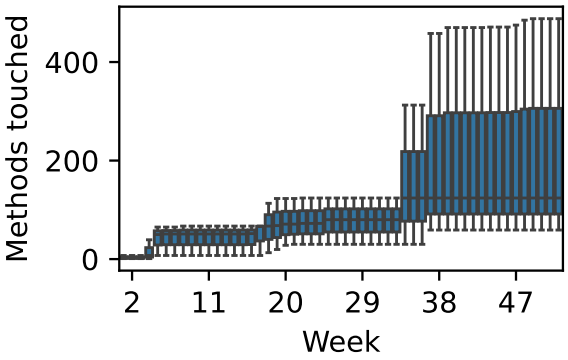
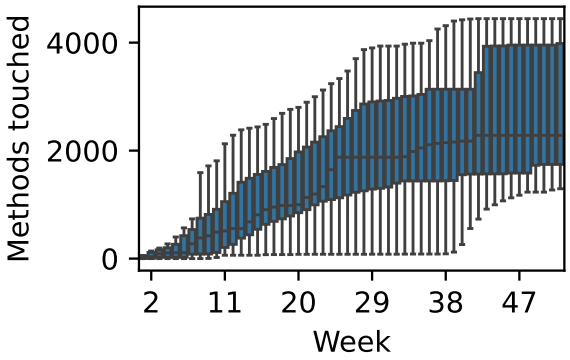


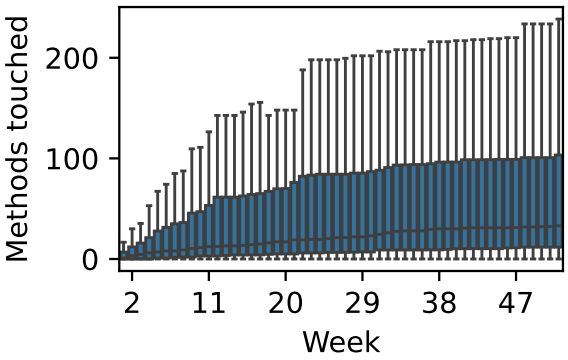
Figure 49: All developers (300) showing methods touched



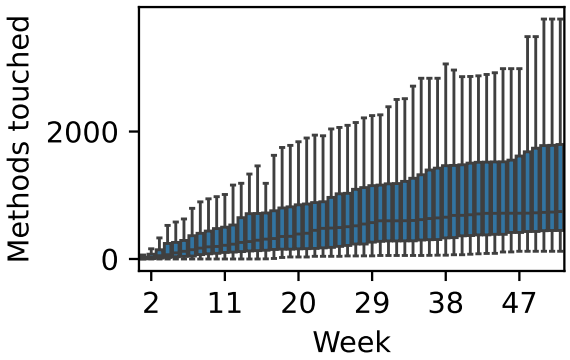
(a) Transient founder developers (3) showing methods touched



(b) Sustained founder developers (7) showing methods touched



(c) Transient later joiner developers (206) showing methods touched



(d) Sustained later joiner developers (84) showing methods touched

Figure 50: A box plot of total methods touched on mean each month, with quartile shading.