Learning a Codebase

Derek Somerville

November 23, 2025

Contents

1 Repository

1.1 Glossary - Summary

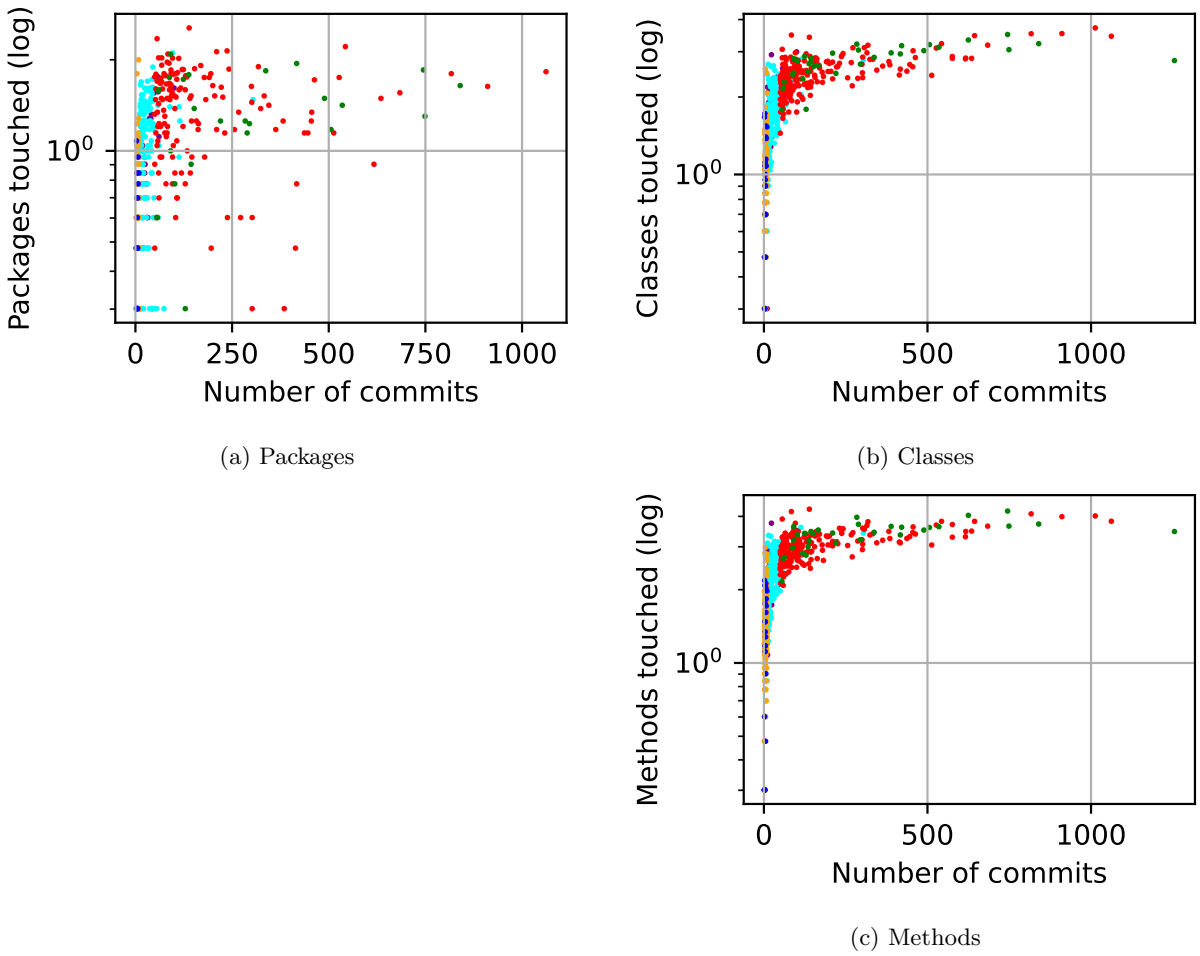
- The founder developer starts in the first six months of a project.
- Late joiner developers begin after six months.
- Transient developers have ten (10) or fewer commits.
- Moderate developers have more than ten (10) commits but fewer than 50 commits or commits for less than 250 days.
- Sustained developers make 50 or more commits and commit for 250 days or more.

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	Transient Founder	Moderate Founder	Sustained Founder	Transient Joiner	Moderate Joiner	Sustained Joiner	Commit	Start	End
3	activiti	4	7	6	47	20	5	2329	2010-Jun-18	On going
5	airbyte-platform	3	2	1	45	32	6	1712	2020-Aug-04	2023-Sep-27
7	ambari	0	3	0	34	32	15	4125	2011-Sep-22	2023-Nov-18
10	automq	2	0	0	40	10	3	616	2011-Sep-07	2017-Apr-21
14	buck	8	5	6	119	85	21	6316	2013-Mar-21	2021-May-17
15	camel	0	4	1	257	63	13	5441	2007-Mar-19	On going
18	checkstyle	0	0	1	57	42	9	2803	2001-Jun-22	On going
20	cxf	6	6	3	27	17	3	1687	2008-Apr-29	On going
21	intellij-community	0	3	7	113	93	84	28741	2004-Nov-11	2018-Apr-18
26	guava	0	1	0	47	8	3	955	2009-Sep-15	2024-Jan-23
28	jenkins	1	1	1	144	39	4	3594	2006-Nov-05	On going
33	openmrs-core	0	1	0	123	26	3	1466	2006-May-03	On going
36	presto	0	1	3	162	71	18	4867	2012-Aug-09	On going
37	quarkus	1	3	5	126	25	8	2682	2018-Jun-22	On going
39	selenium	3	3	0	55	17	8	1718	2004-Nov-03	On going
	Total	28	40	34	1396	580	203	69052		

1.3 Sample scatter touched against commit -



(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 six categories: Transient Founder (Blue, 28), Moderate Founder (Purple, 40), Sustained Founder (Green, 34), Transient Later Joiner (Orange, 203), Moderate Later Joiner (Cyan, 203), Sustained Later Joiner (Red, 203).

1.4 Total commits by developer - Developer Commits

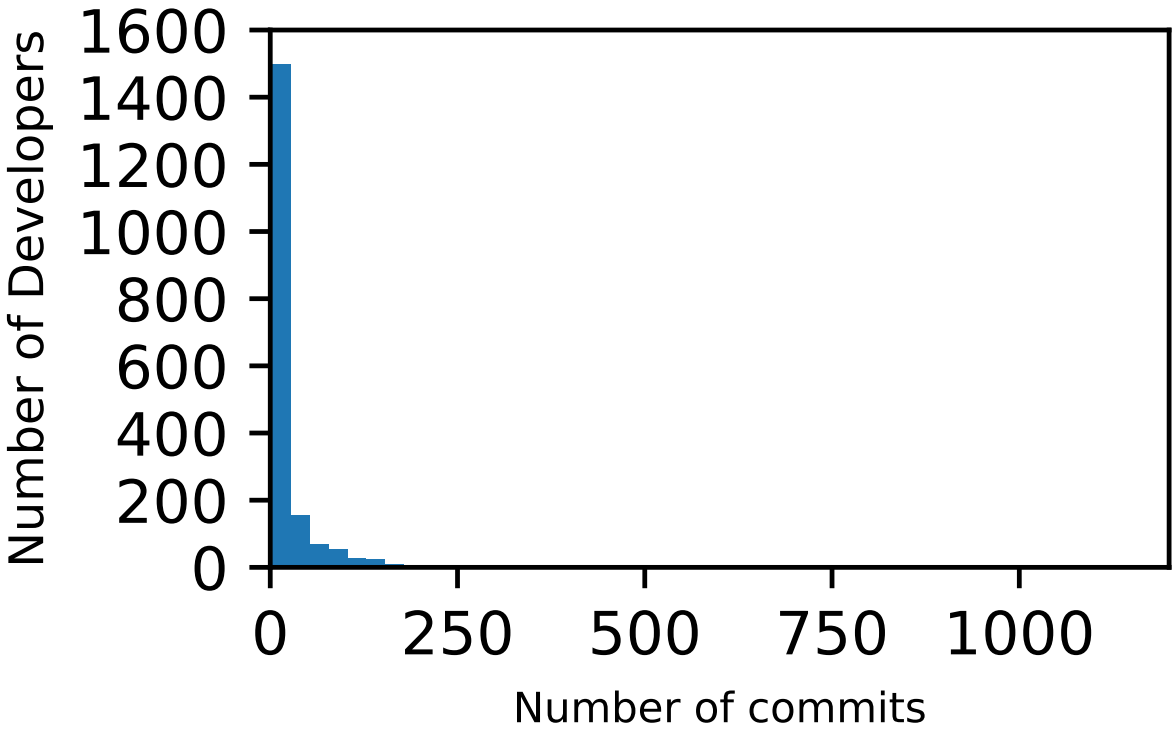
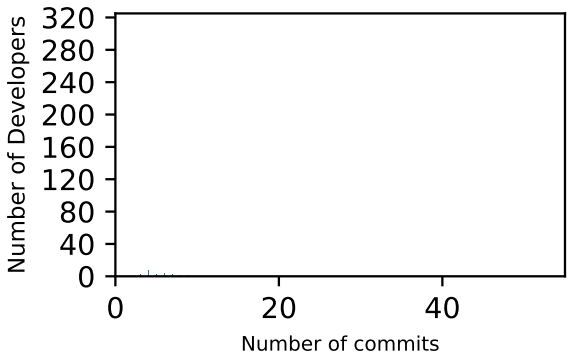
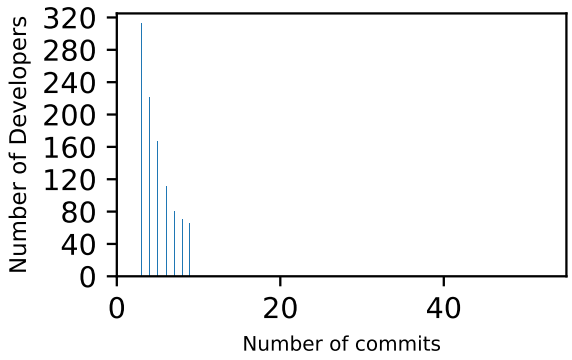


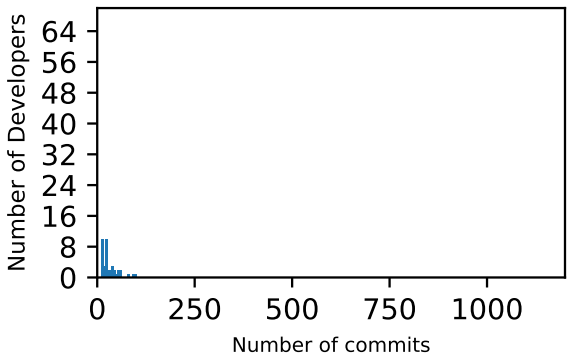
Figure 2: Histogram of number of commits made by all developers 1 in 15 sampled from GitHub, excluding developers with less than three (3). The maximum number of developers is 1497.



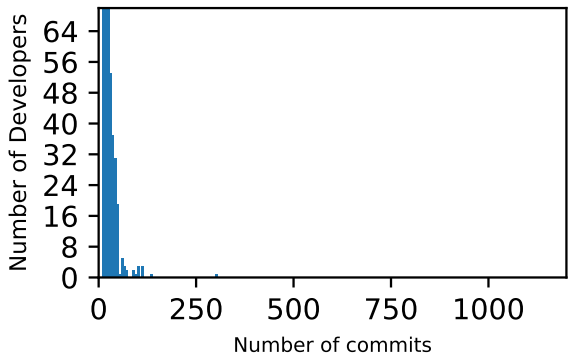
(a) transient founder 19. The maximum number of developers is 7.



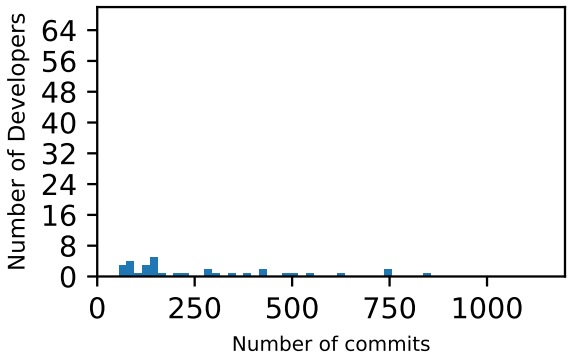
(b) transient later joiner 1030. The maximum number of developers is 313.



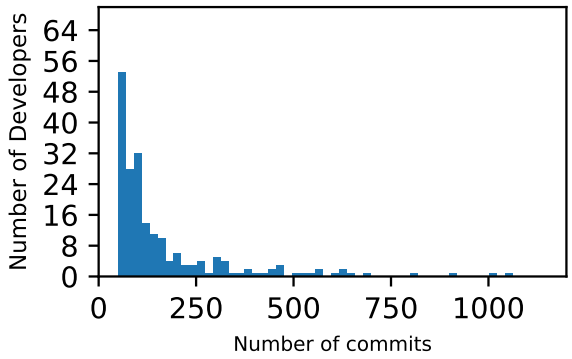
(c) moderate founder 40. The maximum number of developers is 10.



(d) moderate later joiner 580. The maximum number of developers is 216.



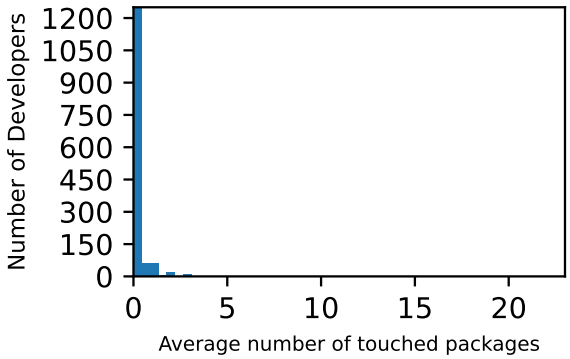
(e) sustained founder 34. The maximum number of developers is 5.



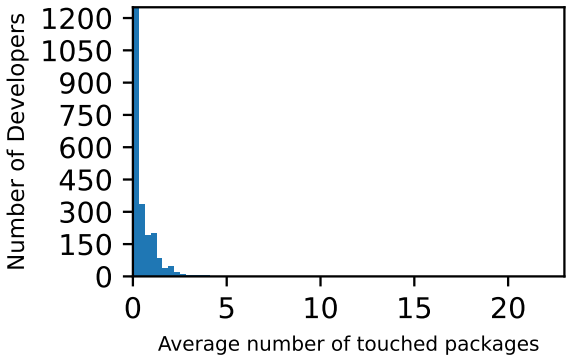
(f) sustained later joiner 203. The maximum number of developers is 53.

Figure 3: Histomagram of number of commits made by six 6 categories of developers in 15 sampled from GitHub, excluding developers with less than three (3)

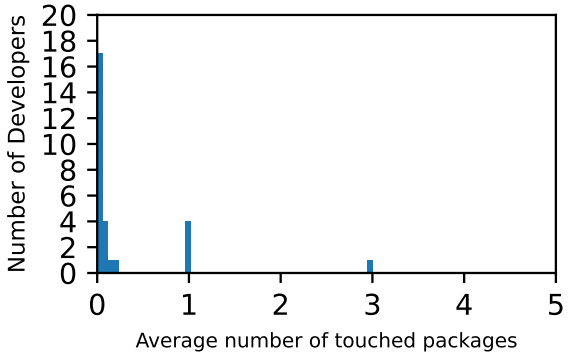
1.5 Histogram components touched by developer - For components with total



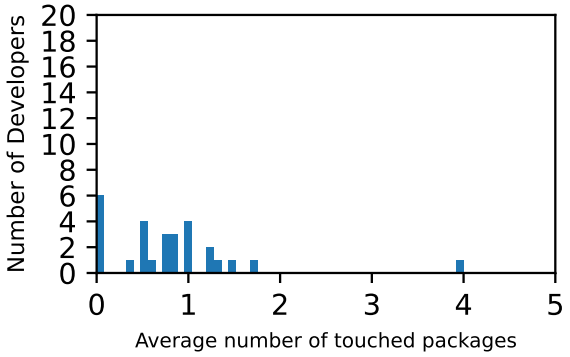
(a) By Day for All 2278 developers. The maximum number of developers is 2108.



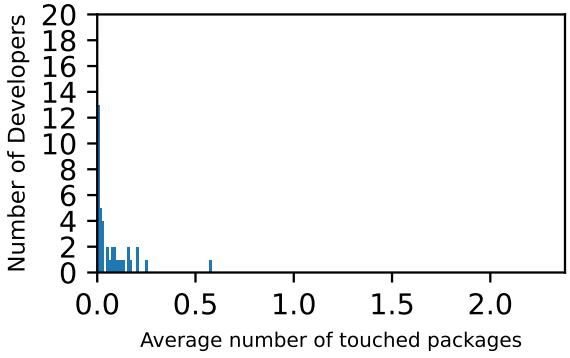
(b) By Commit for All 2278 developers. The maximum number of developers is 1304.



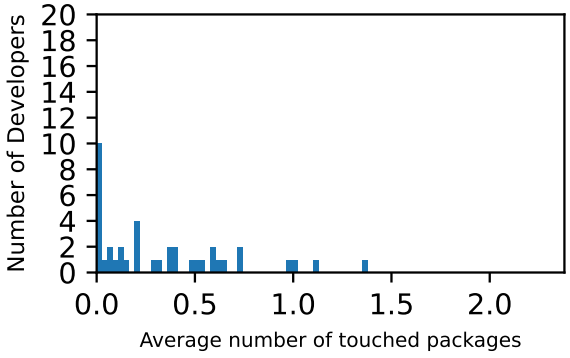
(c) By Day for transient founder 28 developers. The maximum number of developers is 17.



(d) By Commit for transient founder 28 developers. The maximum number of developers is 6.

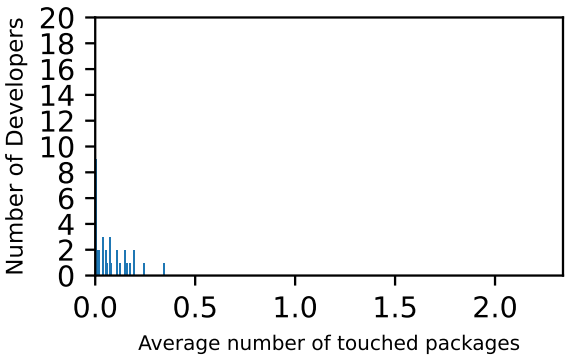


(e) By Day for moderate founder 40 developers. The maximum number of developers is 13.

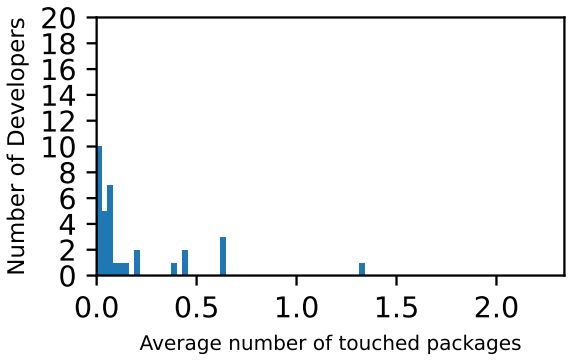


(f) By Commit for moderate founder 40 developers. The maximum number of developers is 10.

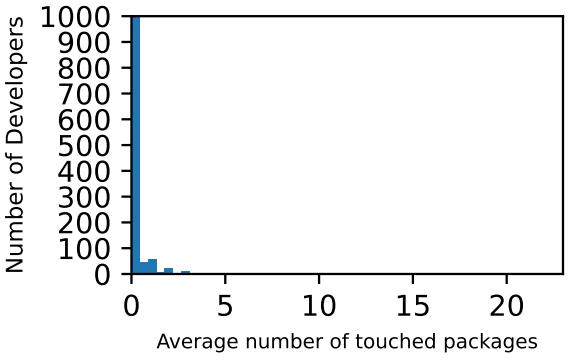
Figure 4: Part 1 continued on next page.



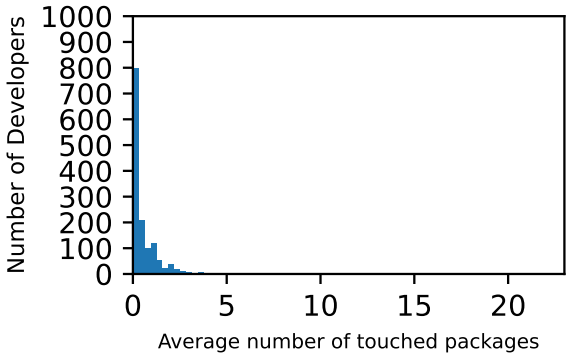
(a) By Day for sustained founder 34 developers. The maximum number of developers is 9.



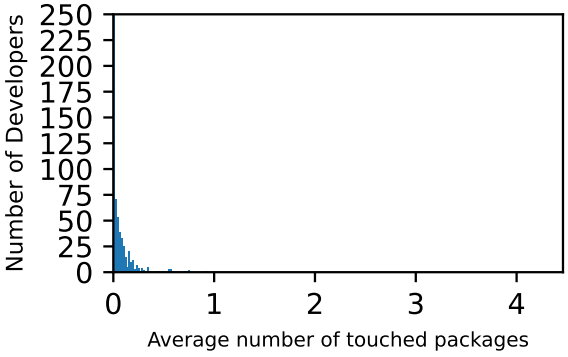
(b) By Commit for sustained founder 34 developers. The maximum number of developers is 10.



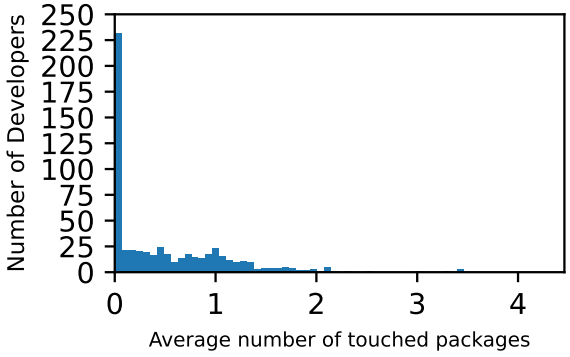
(c) By Day for transient later joiner 1396 developers. The maximum number of developers is 1249.



(d) By Commit for transient later joiner 1396 developers. The maximum number of developers is 798.

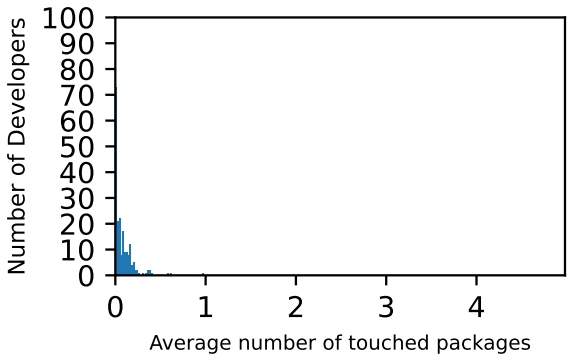


(e) By Day for moderate later joiner 580 developers. The maximum number of developers is 252.

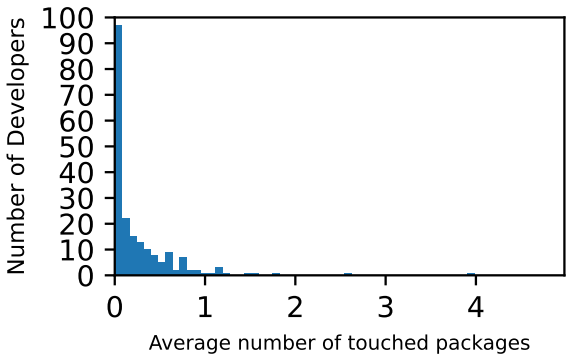


(f) By Commit for moderate later joiner 580 developers. The maximum number of developers is 232.

Figure 5: Part 2 continued on next page.

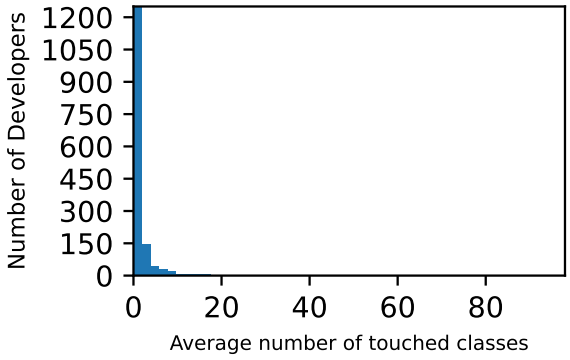


(a) By Day for sustained later joiner 203 developers. The maximum number of developers is 73.

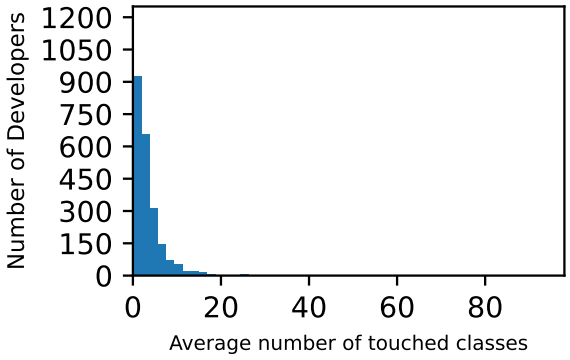


(b) By Commit for sustained later joiner 203 developers. The maximum number of developers is 97.

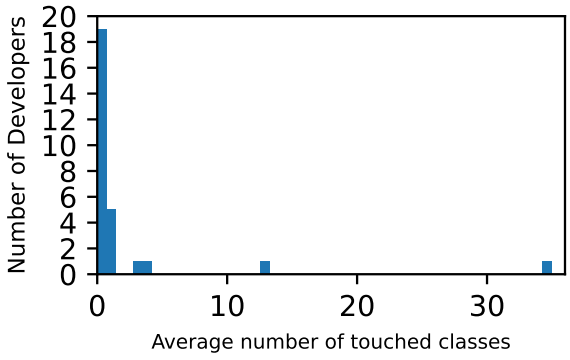
Figure 6: Part 3. Histogram of new packages touched (x-axis) against number of developers (y-axis) from six 6 categories of developers from 15 sampled from GitHub.



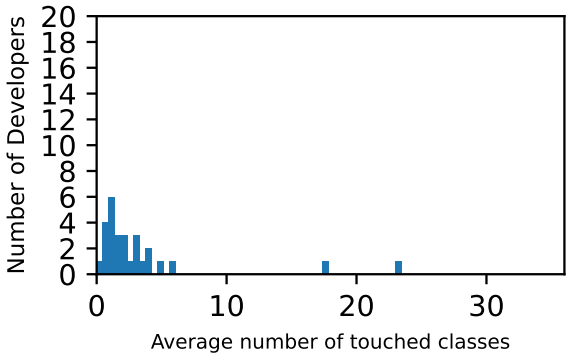
(a) By Day for All 2278 developers. The maximum number of developers is 1999.



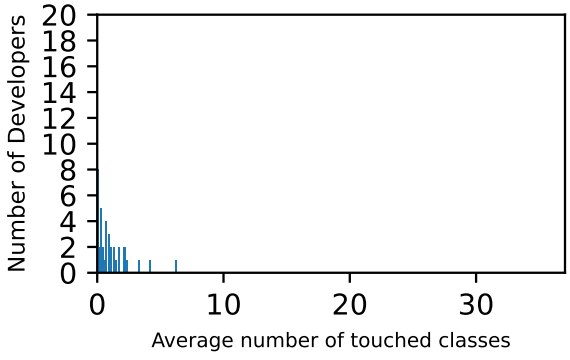
(b) By Commit for All 2278 developers. The maximum number of developers is 926.



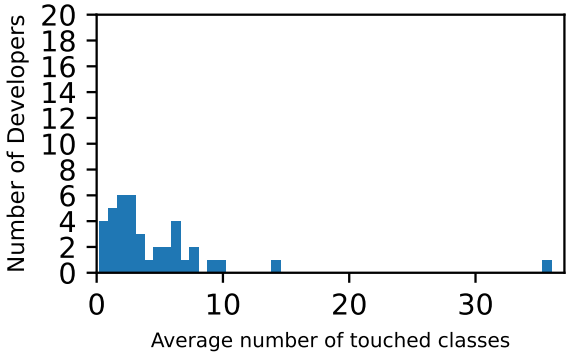
(c) By Day for transient founder 28 developers. The maximum number of developers is 19.



(d) By Commit for transient founder 28 developers. The maximum number of developers is 6.

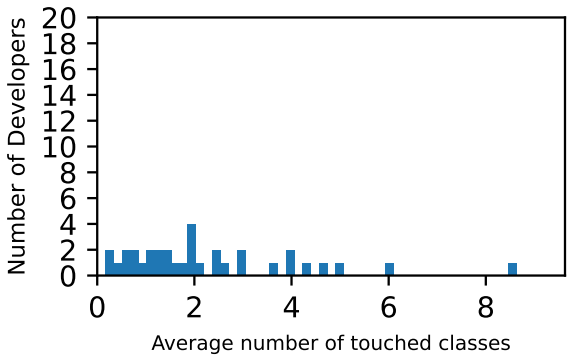


(e) By Day for moderate founder 40 developers. The maximum number of developers is 8.

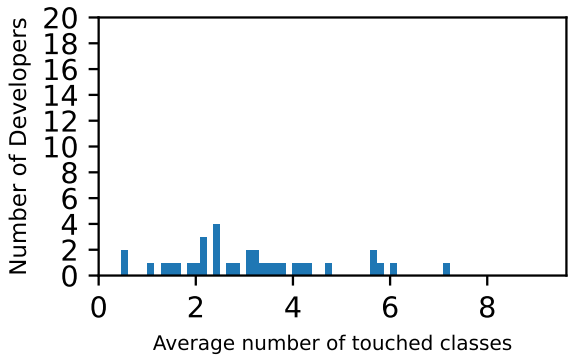


(f) By Commit for moderate founder 40 developers. The maximum number of developers is 6.

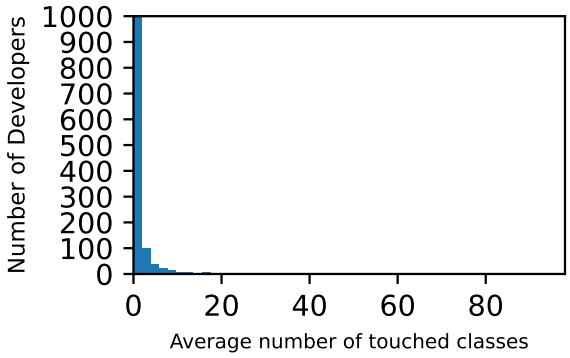
Figure 7: Part 1 continued on next page.



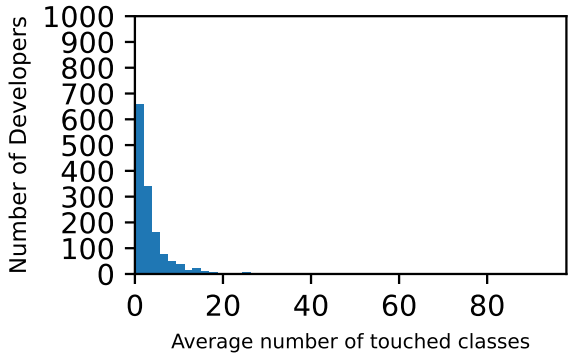
(a) By Day for sustained founder 34 developers. The maximum number of developers is 4.



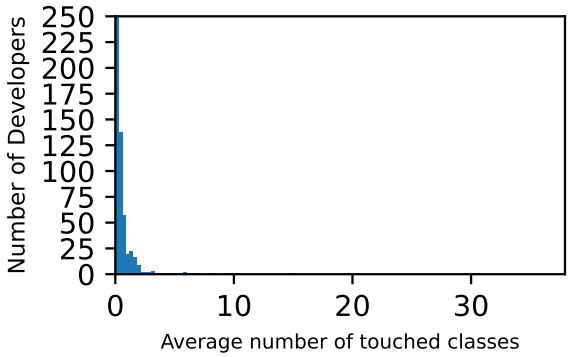
(b) By Commit for sustained founder 34 developers. The maximum number of developers is 4.



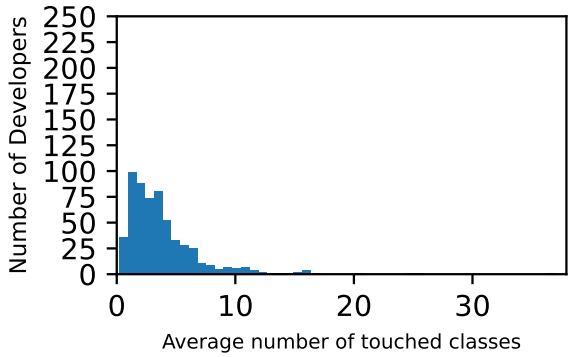
(c) By Day for transient later joiner 1396 developers. The maximum number of developers is 1195.



(d) By Commit for transient later joiner 1396 developers. The maximum number of developers is 657.

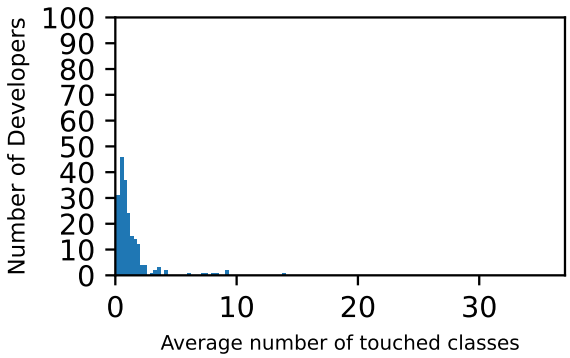


(e) By Day for moderate later joiner 580 developers. The maximum number of developers is 299.

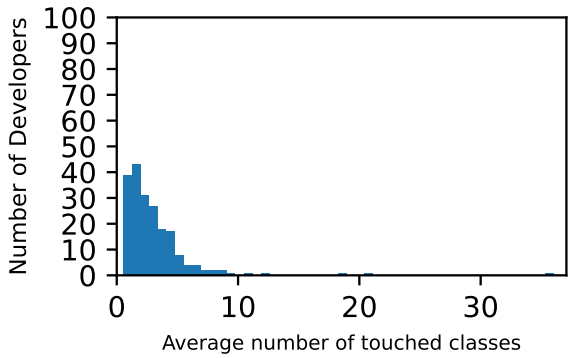


(f) By Commit for moderate later joiner 580 developers. The maximum number of developers is 99.

Figure 8: Part 2 continued on next page.

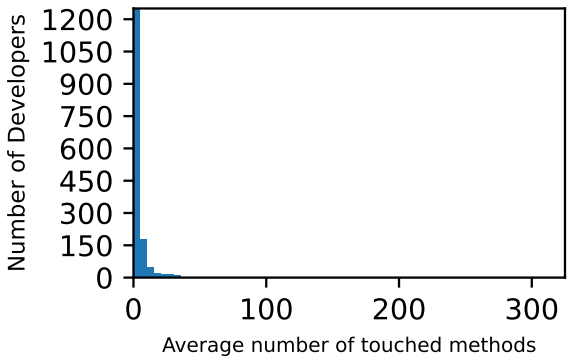


(a) By Day for sustained later joiner 203 developers. The maximum number of developers is 46.

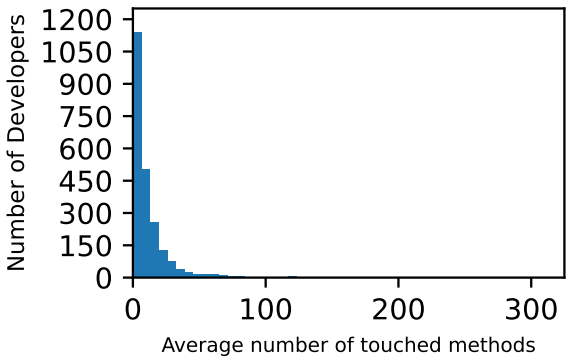


(b) By Commit for sustained later joiner 203 developers. The maximum number of developers is 43.

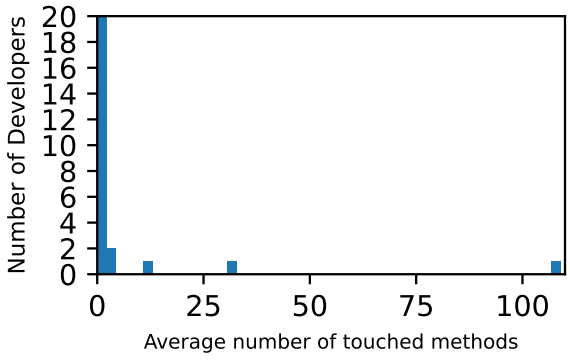
Figure 9: Part 3. Histogram of new classes touched (x-axis) against number of developers (y-axis) from six 6 categories of developers from 15 sampled from GitHub.



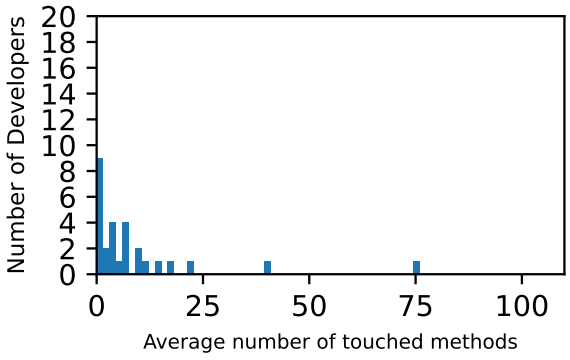
(a) By Day for All 2278 developers. The maximum number of developers is 1955.



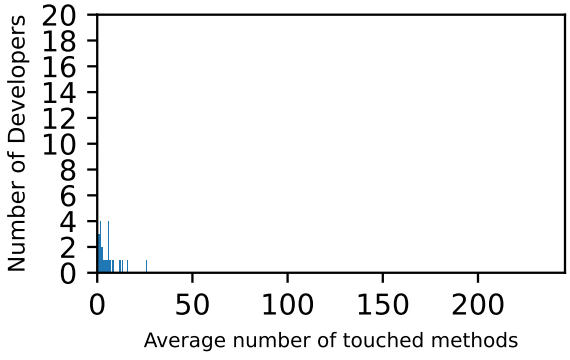
(b) By Commit for All 2278 developers. The maximum number of developers is 1138.



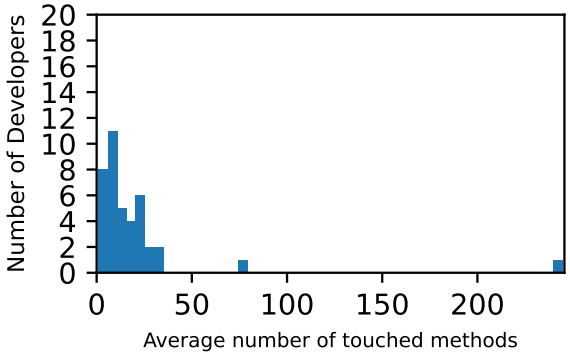
(c) By Day for transient founder 28 developers. The maximum number of developers is 23.



(d) By Commit for transient founder 28 developers. The maximum number of developers is 9.

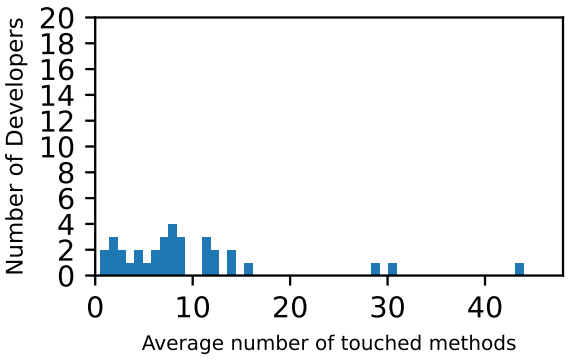


(e) By Day for moderate founder 40 developers. The maximum number of developers is 8.

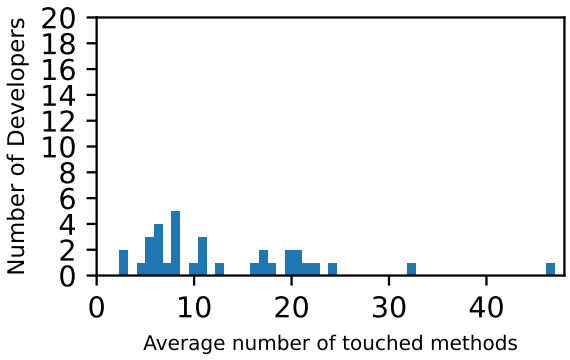


(f) By Commit for moderate founder 40 developers. The maximum number of developers is 11.

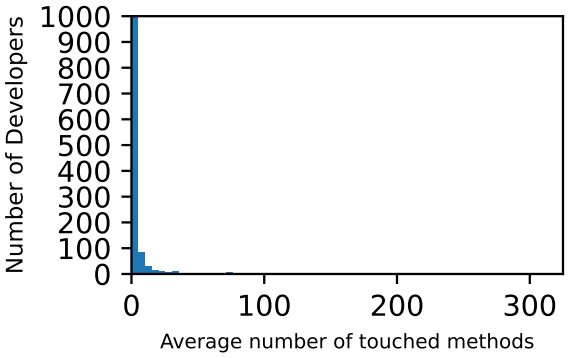
Figure 10: Part 1 continued on next page.



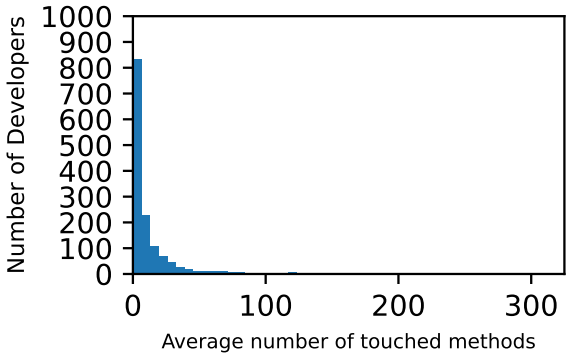
(a) By Day for sustained founder 34 developers. The maximum number of developers is 4.



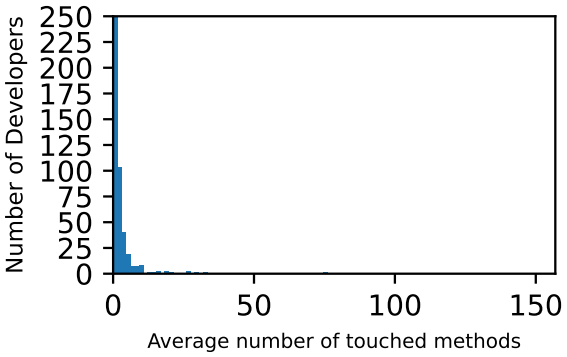
(b) By Commit for sustained founder 34 developers. The maximum number of developers is 5.



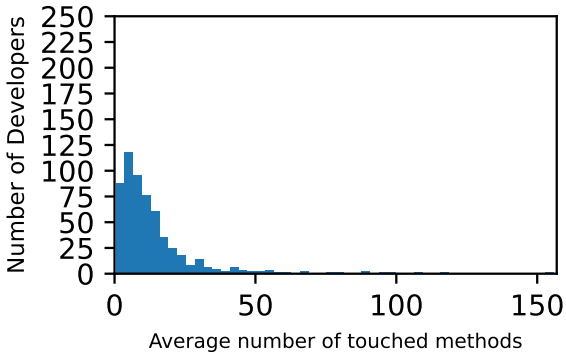
(c) By Day for transient later joiner 1396 developers. The maximum number of developers is 1218.



(d) By Commit for transient later joiner 1396 developers. The maximum number of developers is 831.

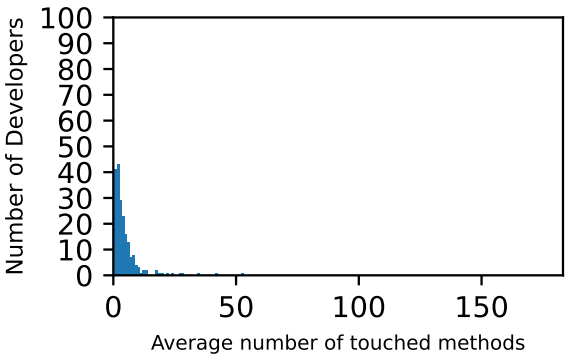


(e) By Day for moderate later joiner 580 developers. The maximum number of developers is 384.

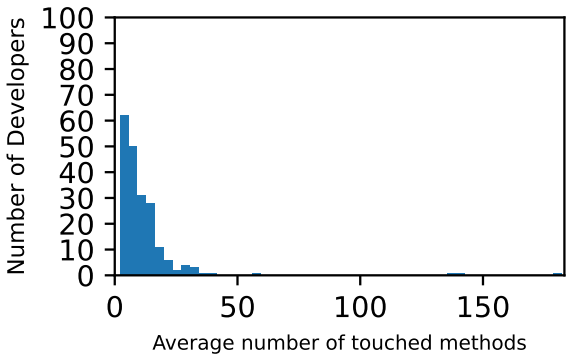


(f) By Commit for moderate later joiner 580 developers. The maximum number of developers is 118.

Figure 11: Part 2 continued on next page.



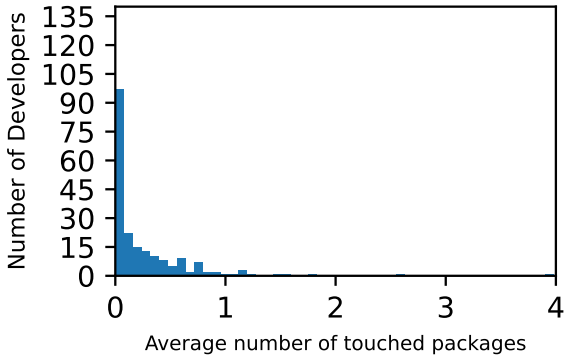
(a) By Day for sustained later joiner 203 developers. The maximum number of developers is 43.



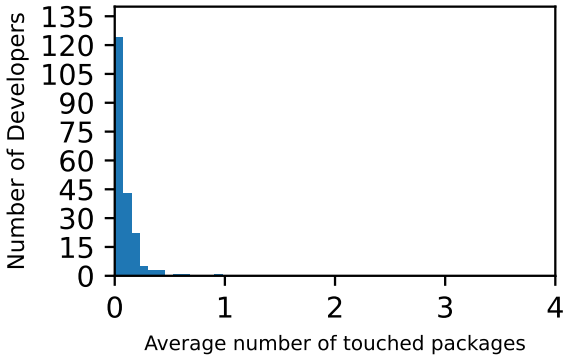
(b) By Commit for sustained later joiner 203 developers. The maximum number of developers is 62.

Figure 12: Part 3. Histogram of new methods touched (x-axis) against number of developers (y-axis) from six 6 categories of developers from 15 sampled from GitHub.

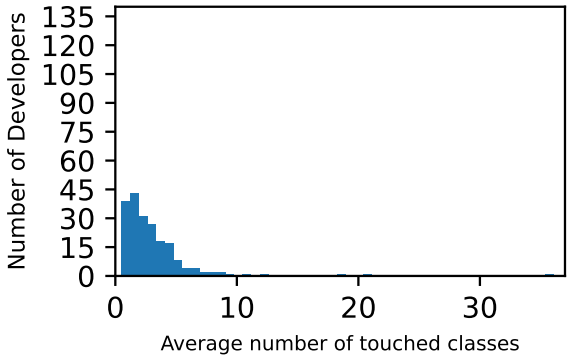
1.6 Average components touched by developer - By commit and by day



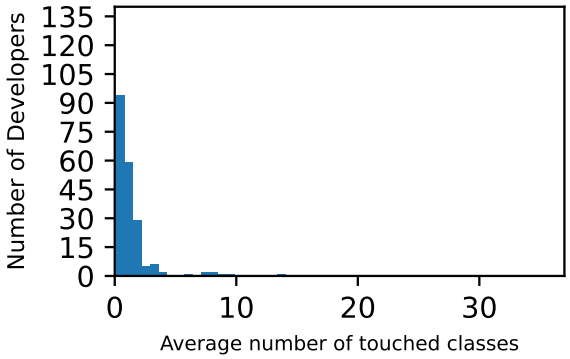
(a) Commit new packages. The maximum number of developers is 97.



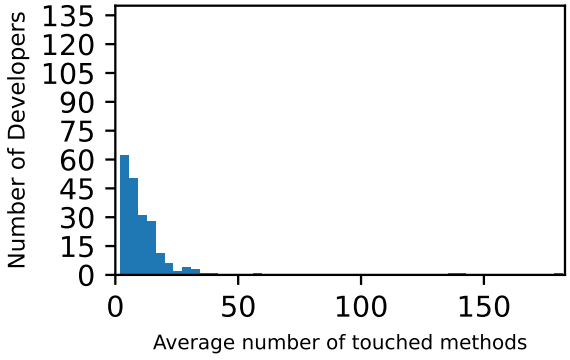
(b) Daily new packages. The maximum number of developers is 124.



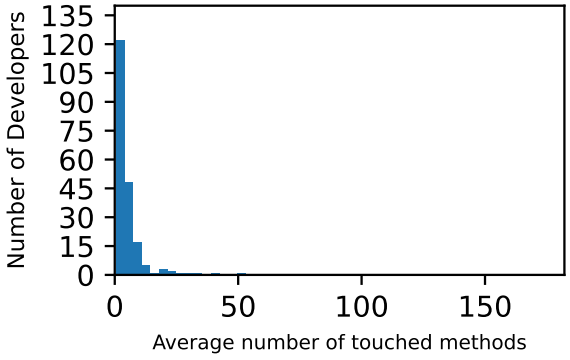
(c) Commit new classes. The maximum number of developers is 43.



(d) Daily new classes. The maximum number of developers is 94.



(e) Commit new methods. The maximum number of developers is 62.



(f) Daily new methods. The maximum number of developers is 122.

Figure 13: 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.

2 Repository: 14

2.1 Time series components touched - 14 For packages touched for each period

A time series of packages touched on average each month.

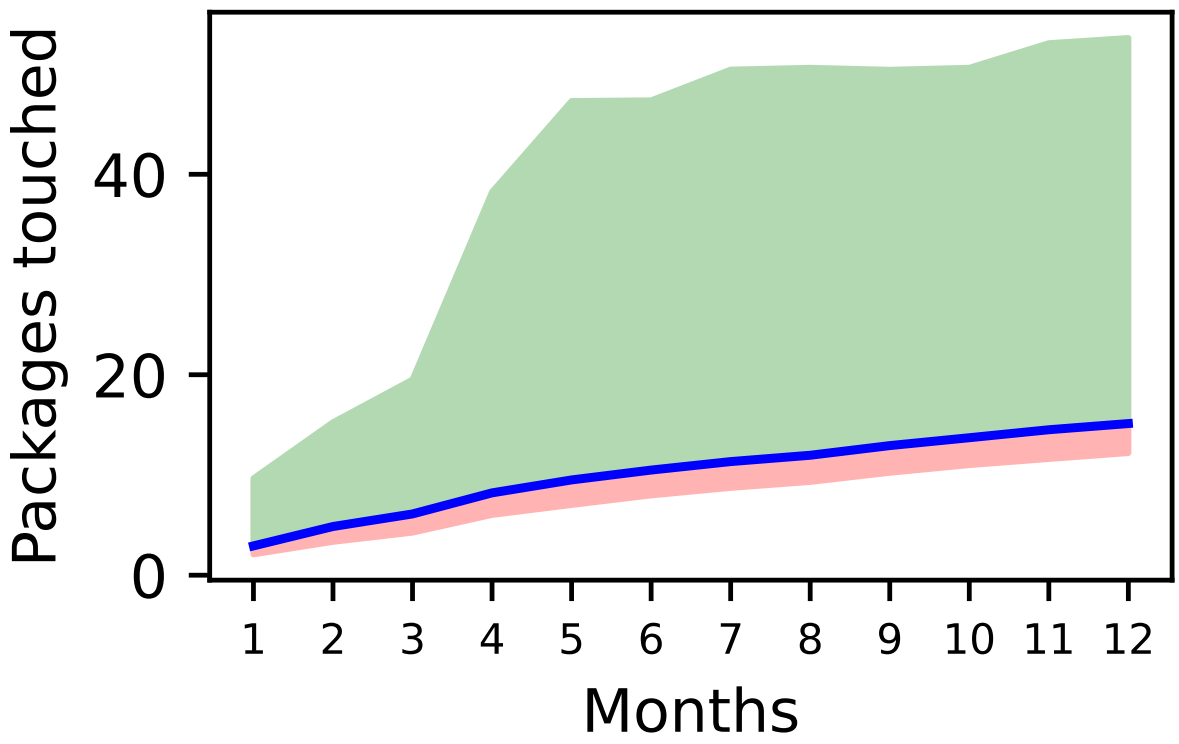
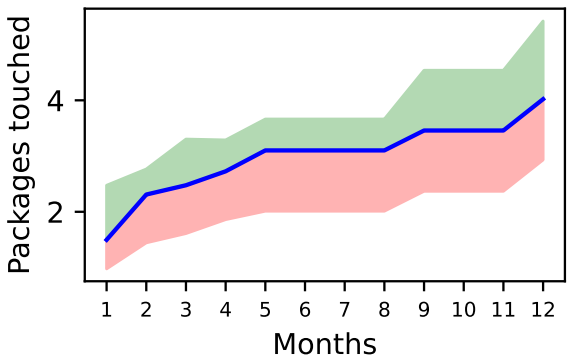
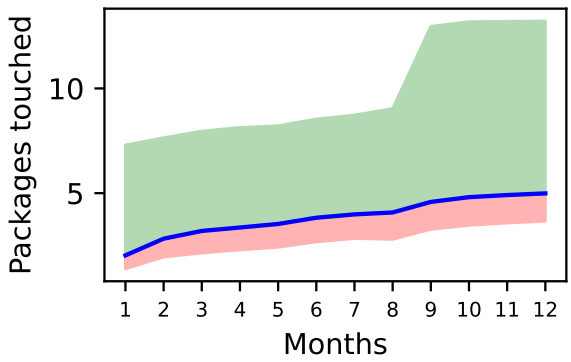


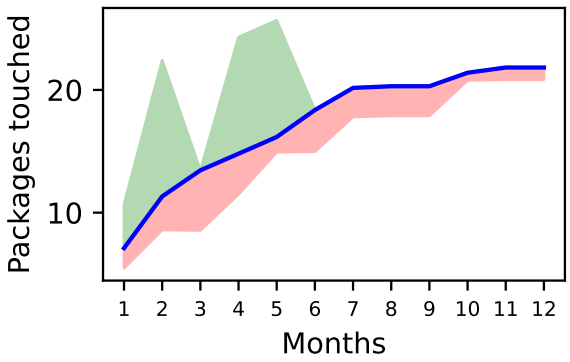
Figure 14: A time series of the number of months (x-axis) against the average (mean) total packages touched (y-axis) for all developers 154,



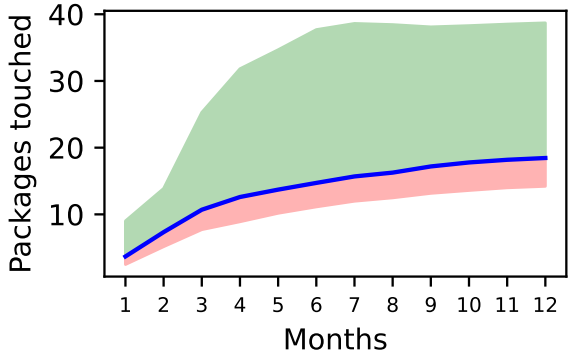
(a) Transient founder developers (8) showing packages touched



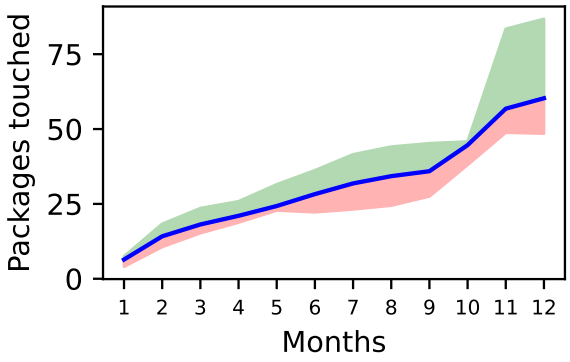
(b) Transient later joiner developers (119) showing packages touched



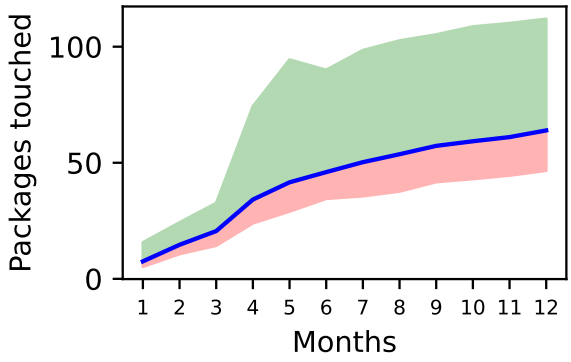
(c) Moderate founder developers (5) showing packages touched



(d) Moderate later joiner developers (85) showing packages touched



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



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

Figure 15: A time series of the number of month (x-axis) against the average (mean) total packages touched (y-axis) for six categories of developer,

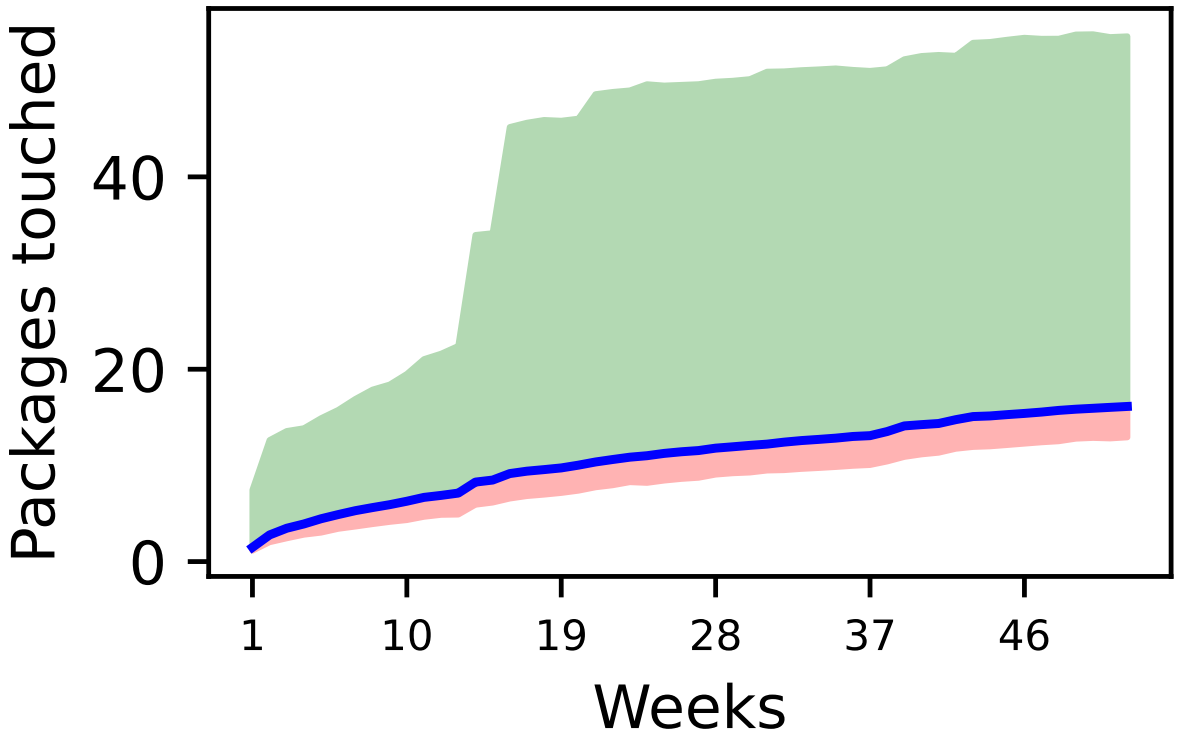
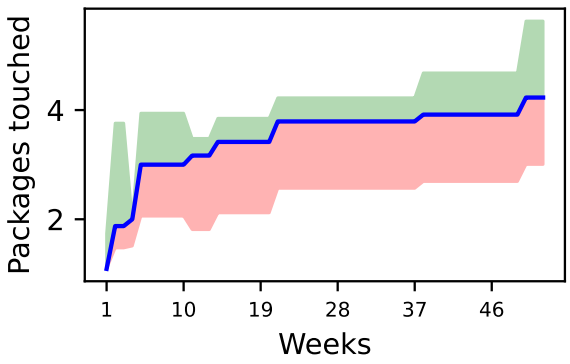
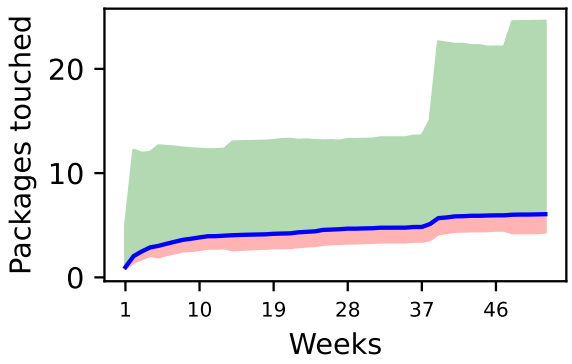


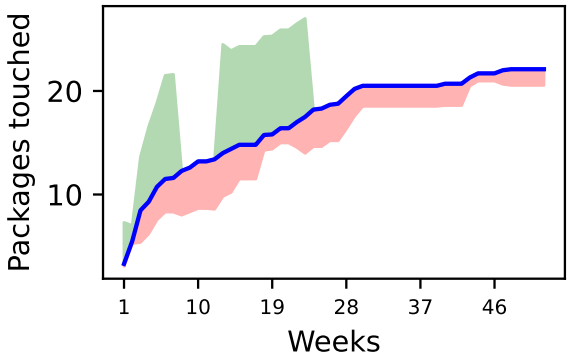
Figure 16: A time series of the number of weeks (x-axis) against the average (mean) total packages touched (y-axis) for all developers 154,



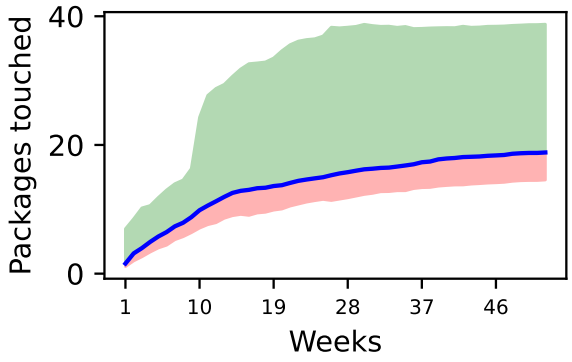
(a) Transient founder developers (8) showing packages touched



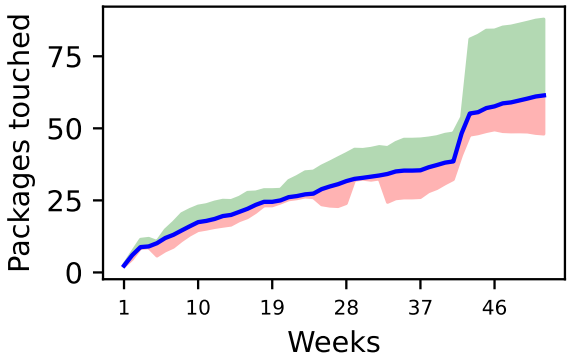
(b) Transient later joiner developers (119) showing packages touched



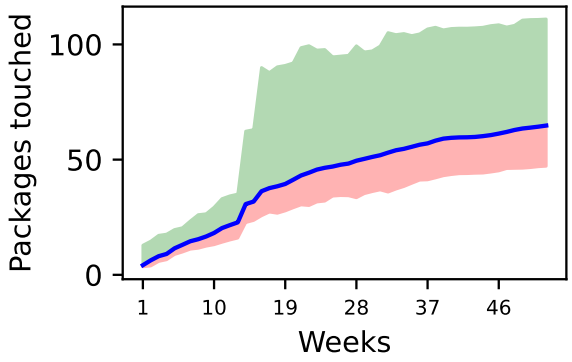
(c) Moderate founder developers (5) showing packages touched



(d) Moderate later joiner developers (85) showing packages touched



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



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

Figure 17: A time series of the number of week (x-axis) against the average (mean) total packages touched (y-axis) for six categories of developer,

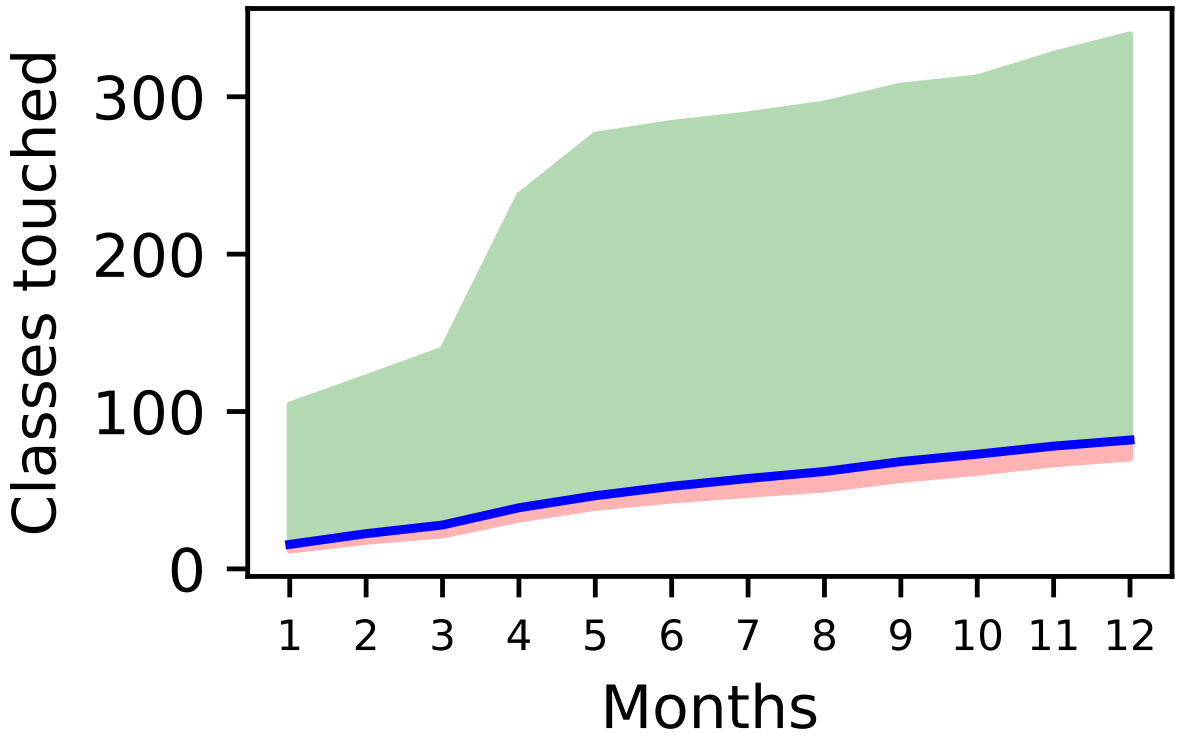
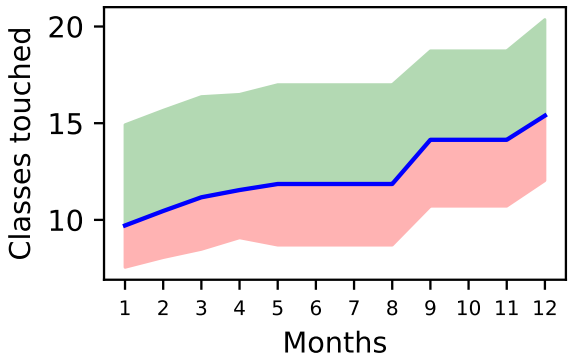
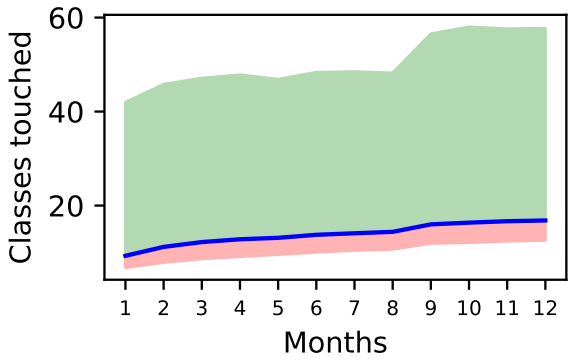


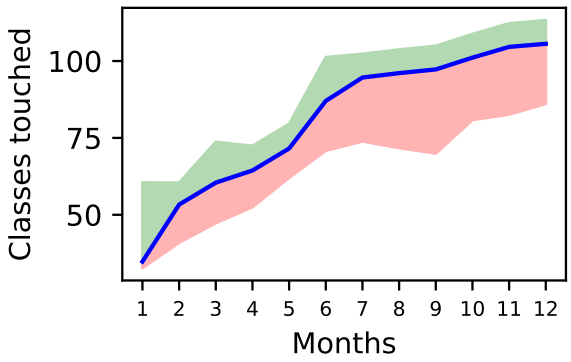
Figure 18: A time series of the number of months (x-axis) against the average (mean) total classes touched (y-axis) for all developers 154,



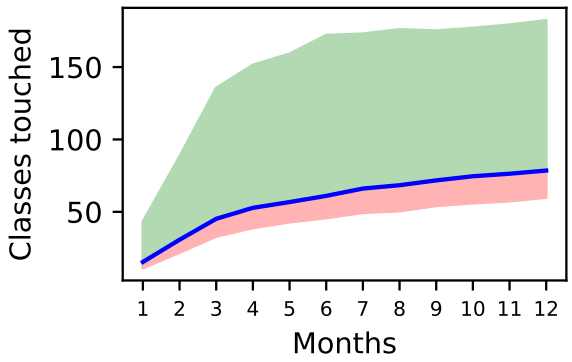
(a) Transient founder developers (8) showing classes touched



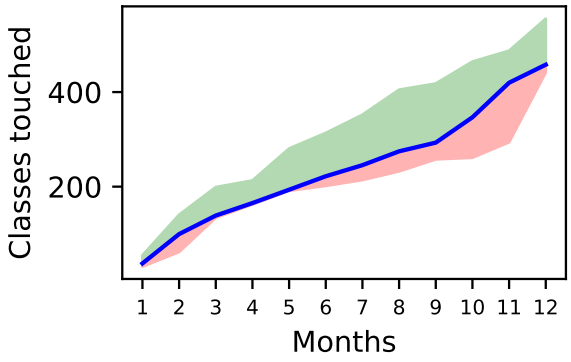
(b) Transient later joiner developers (119) showing classes touched



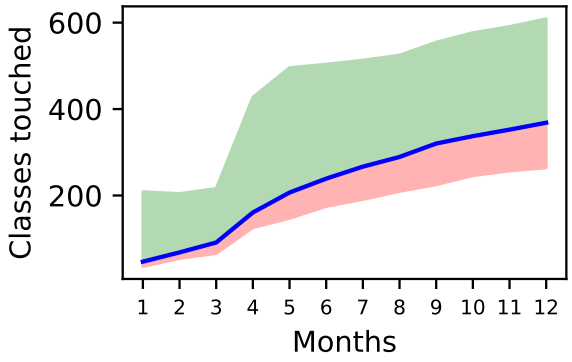
(c) Moderate founder developers (5) showing classes touched



(d) Moderate later joiner developers (85) showing classes touched



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



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

Figure 19: A time series of the number of month (x-axis) against the average (mean) total classes touched (y-axis) for six categories of developer,

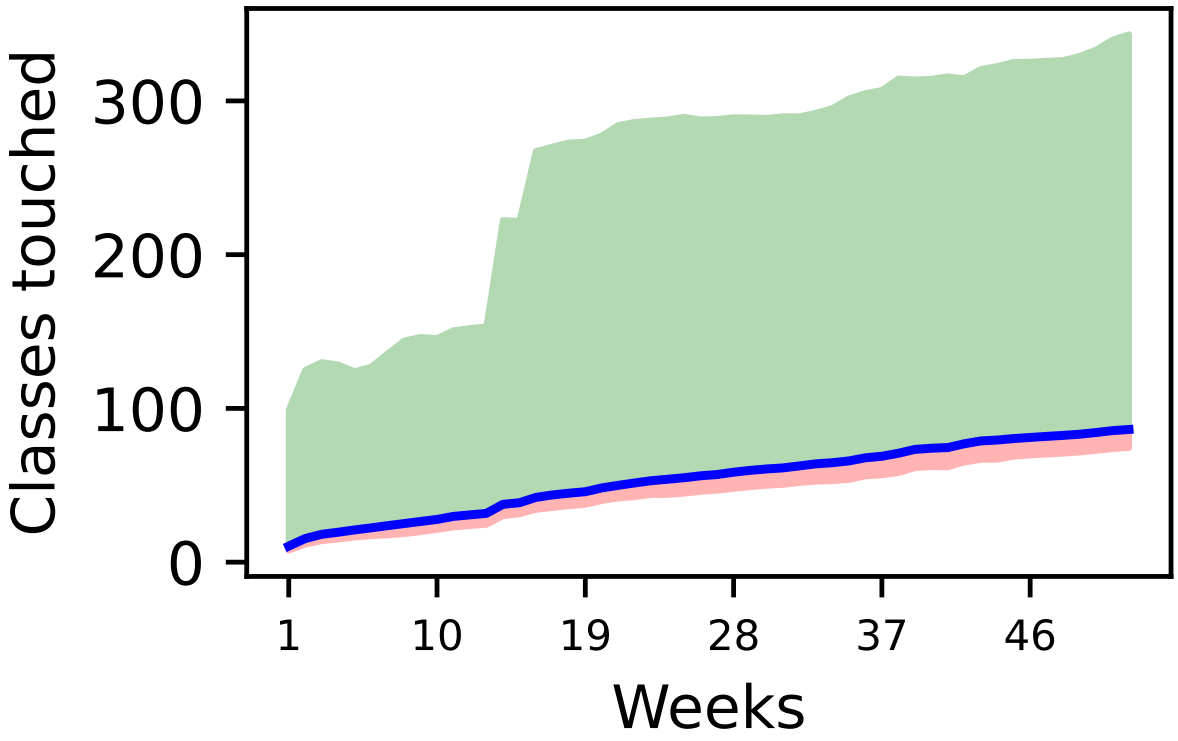
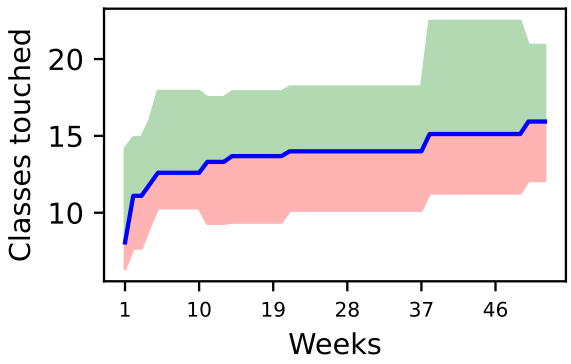
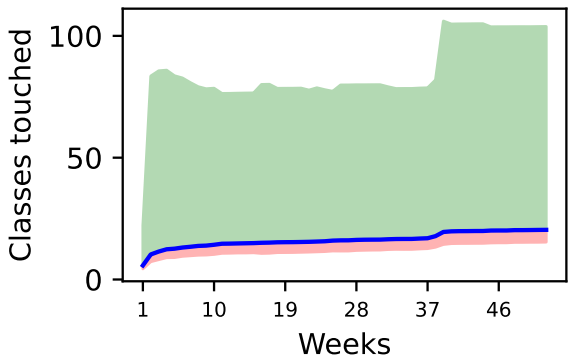


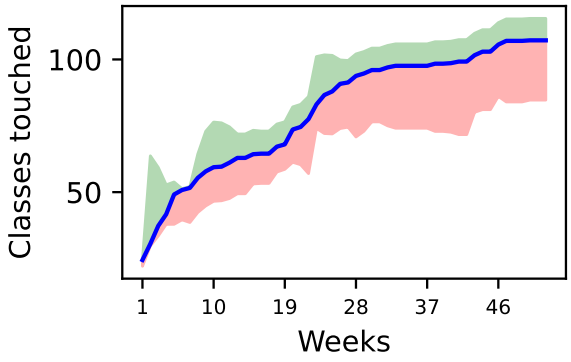
Figure 20: A time series of the number of weeks (x-axis) against the average (mean) total classes touched (y-axis) for all developers 154,



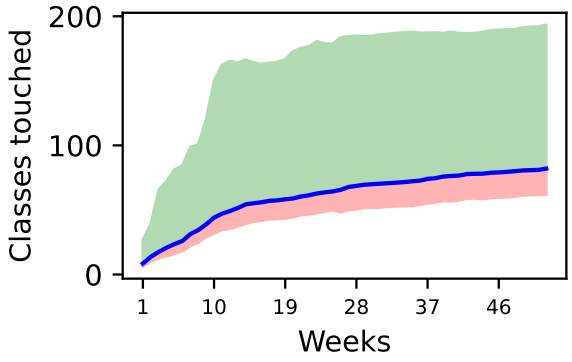
(a) Transient founder developers (8) showing classes touched



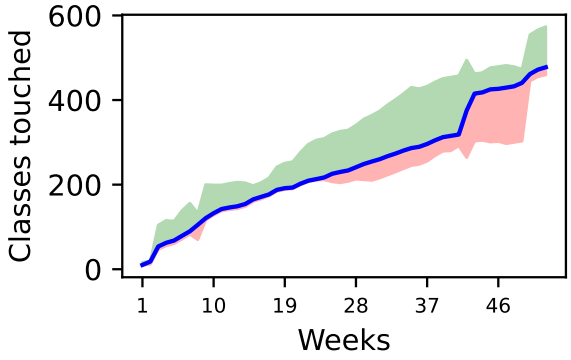
(b) Transient later joiner developers (119) showing classes touched



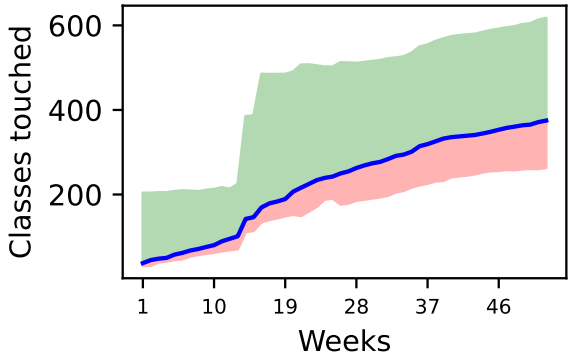
(c) Moderate founder developers (5) showing classes touched



(d) Moderate later joiner developers (85) showing classes touched



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



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

Figure 21: A time series of the number of week (x-axis) against the average (mean) total classes touched (y-axis) for six categories of developer,

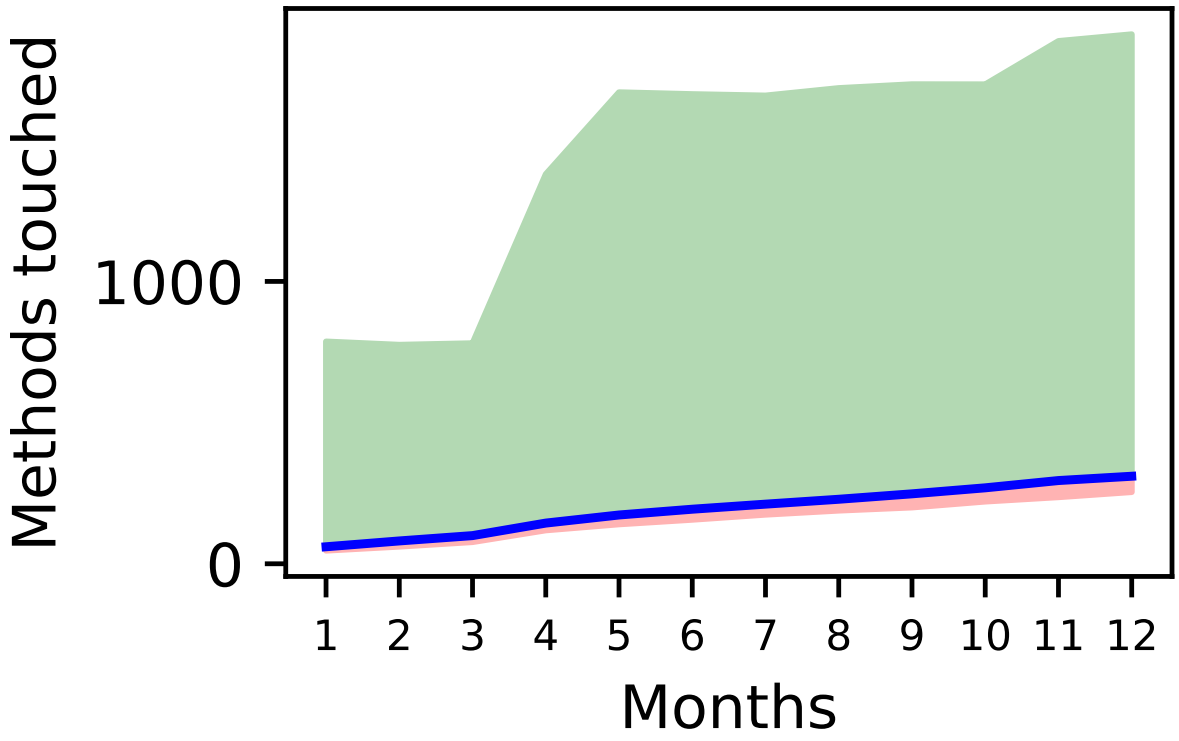
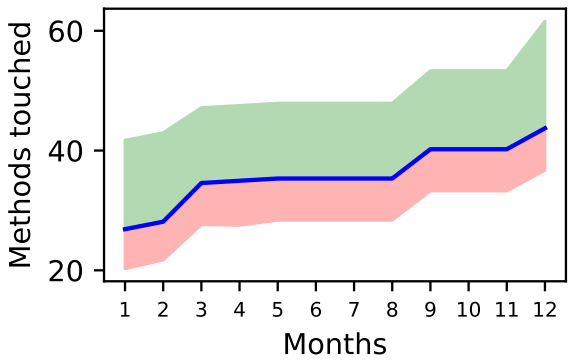
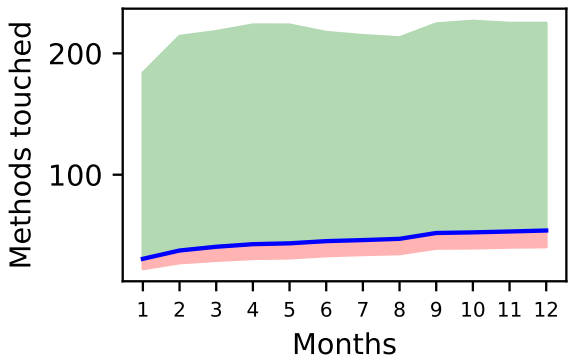


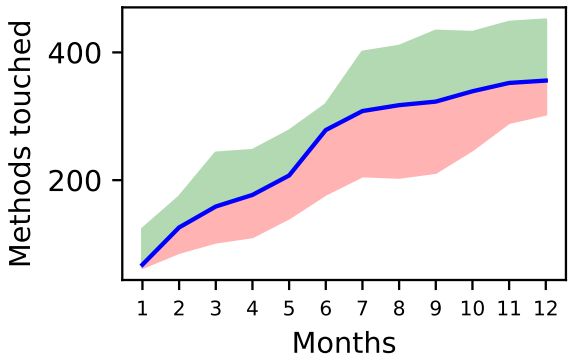
Figure 22: A time series of the number of months (x-axis) against the average (mean) total methods touched (y-axis) for all developers 154,



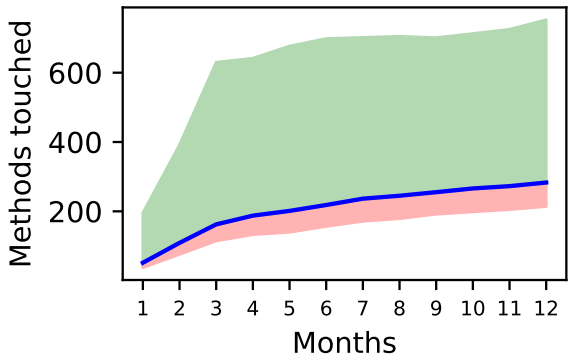
(a) Transient founder developers (8) showing methods touched



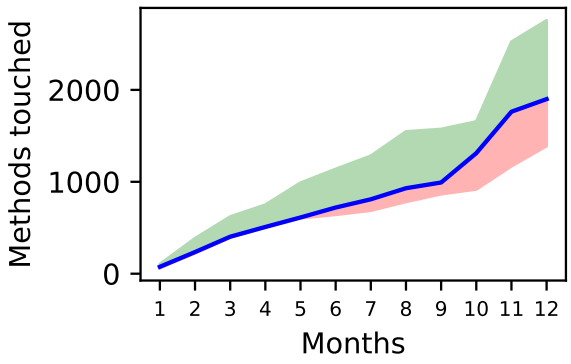
(b) Transient later joiner developers (119) showing methods touched



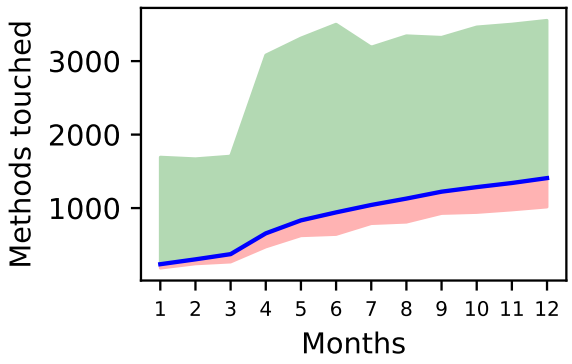
(c) Moderate founder developers (5) showing methods touched



(d) Moderate later joiner developers (85) showing methods touched



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



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

Figure 23: A time series of the number of month (x-axis) against the average (mean) total methods touched (y-axis) for six categories of developer,

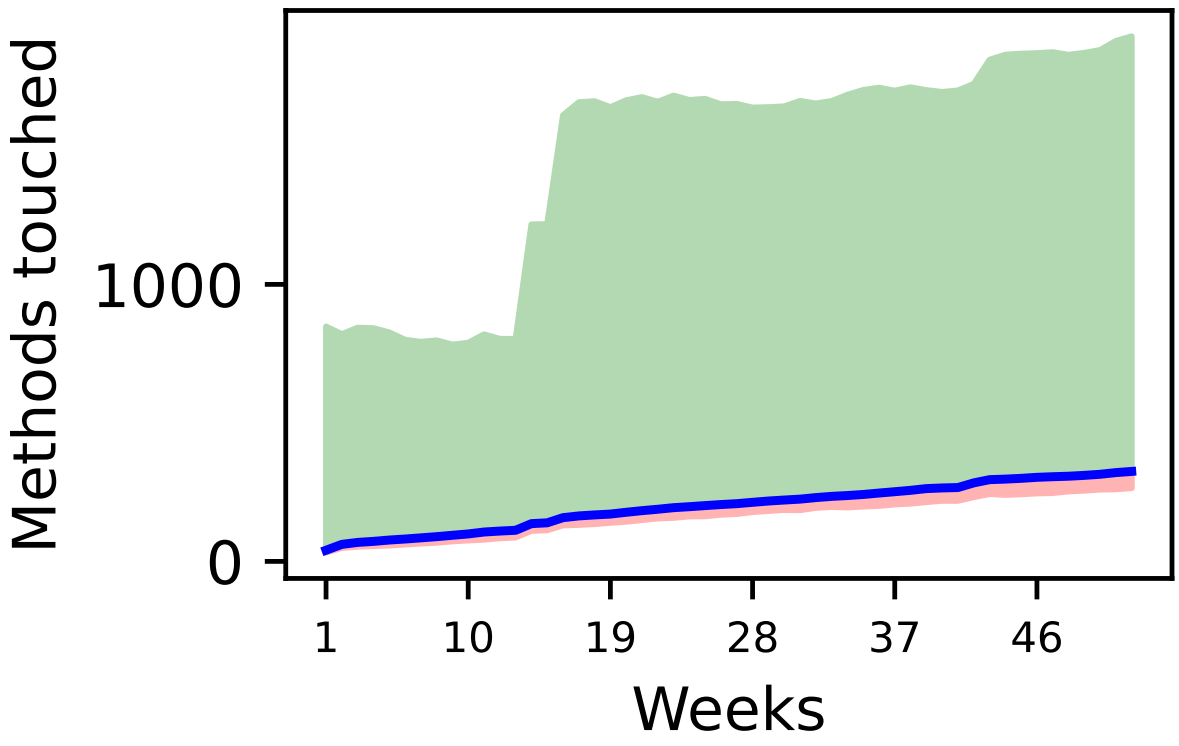
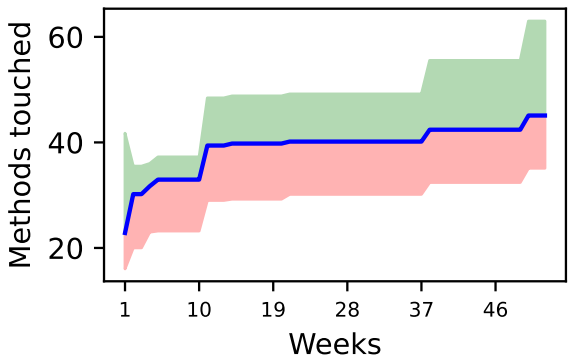
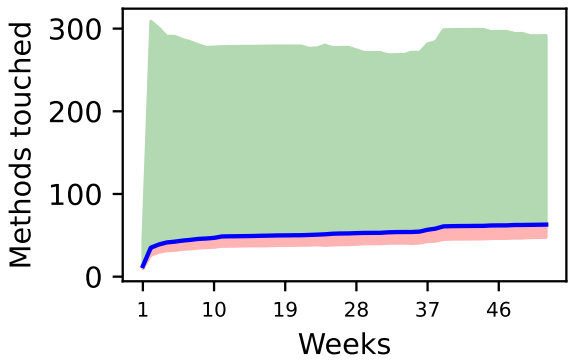


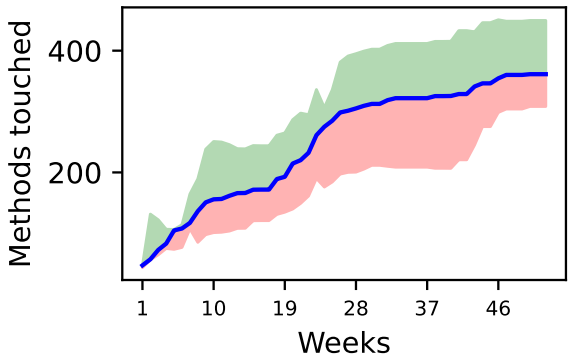
Figure 24: A time series of the number of weeks (x-axis) against the average (mean) total methods touched (y-axis) for all developers 154,



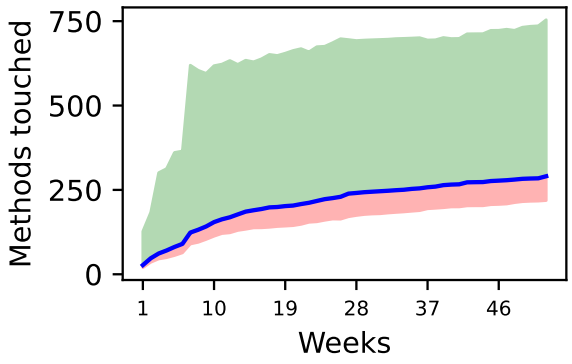
(a) Transient founder developers (8) showing methods touched



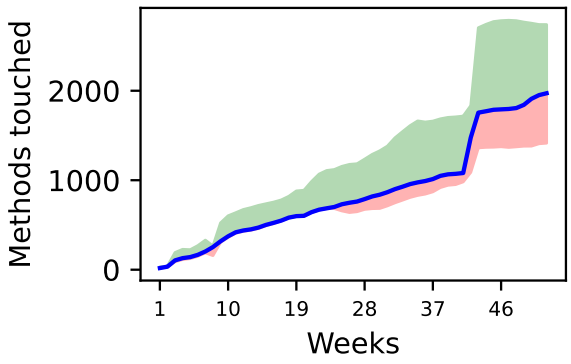
(b) Transient later joiner developers (119) showing methods touched



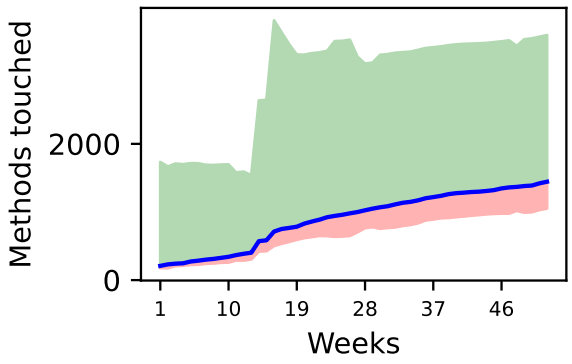
(c) Moderate founder developers (5) showing methods touched



(d) Moderate later joiner developers (85) showing methods touched



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



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

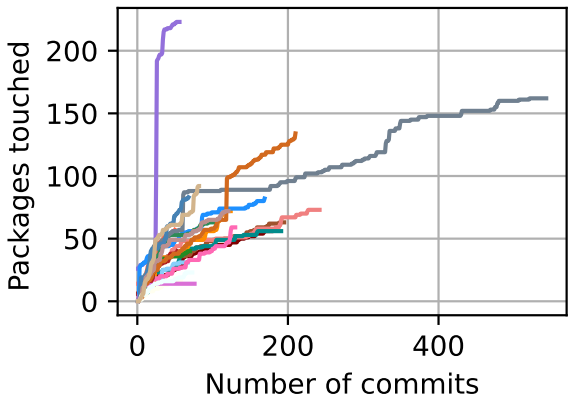
Figure 25: A time series of the number of week (x-axis) against the average (mean) total methods touched (y-axis) for six categories of developer,

2.2 Scatter components touched by developer - 14 For developer commits and components touched.

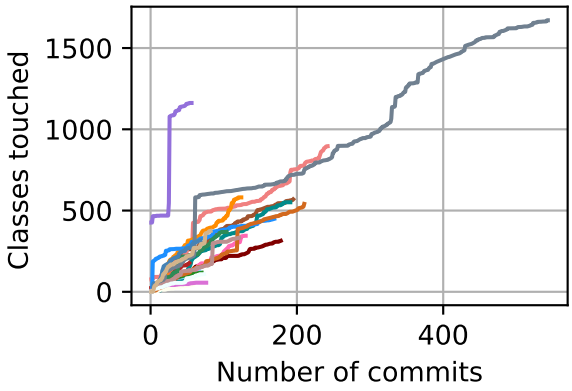
A scatter plot for the first 21 sustained late joiner developers showing the number of commits to components touched.

Table 2: Table of first 21 developers in the Scatter Developer graphs.

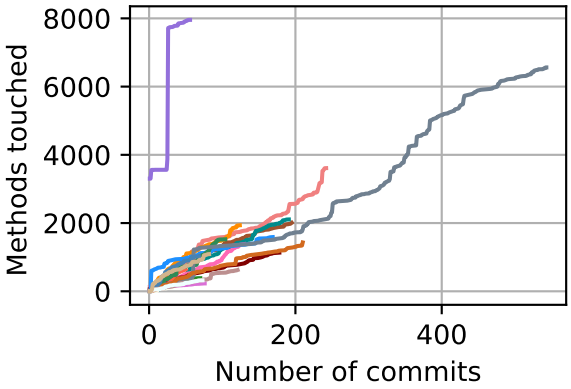
Developer ID	Colour
1774	Maroon
2986	Sienna
3000	LightCoral
3020	DarkCyan
3037	DarkOrange
3050	MediumPurple
3071	LightSkyBlue
3072	HotPink
3141	ForestGreen
3143	FireBrick
3163	DodgerBlue
3172	Tomato
3178	SeaGreen
3216	SlateGray
3217	Azure
3272	Orchid
3339	Chocolate
3354	SteelBlue
3456	RosyBrown
3473	Azure
3485	Tan



(a) packages



(b) classes

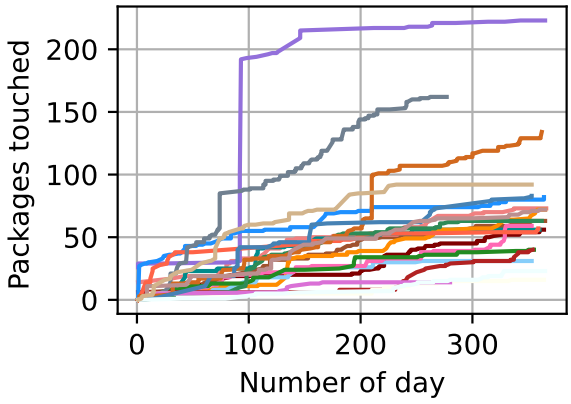


(c) methods

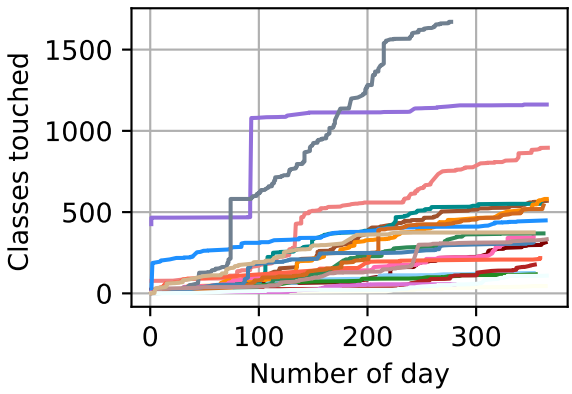
Figure 26: First twenty-one 21 sustained later joiner developers from this repository. The number of commits against the number of components touched. Developers with colours are shown in Table 3.

2.3 Scatter components touched by developer - 14 For developer day and components touched.

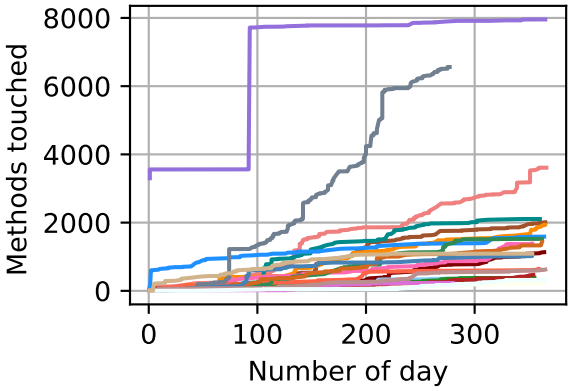
A scatter plot for the first 21 sustained late joiner developers showing the number of day to components touched.



(a) packages



(b) classes



(c) methods

Figure 27: First twenty-one 21 sustained later joiner developers from this repository. The number of day against the number of components touched. Developers with colours are shown in Table 3.

3 Repository: 21

3.1 Time series components touched - 21 For packages touched for each period

A time series of packages touched on average each month.

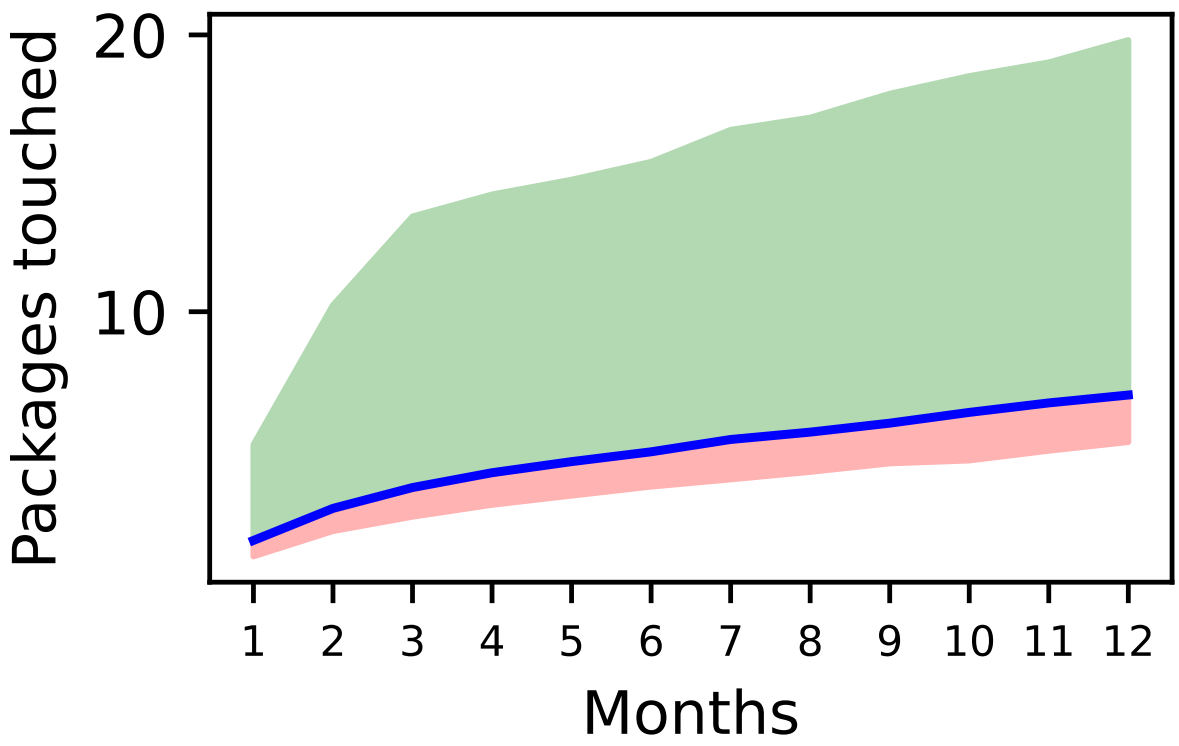
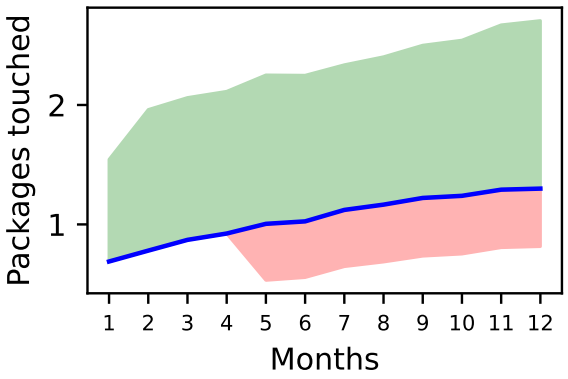
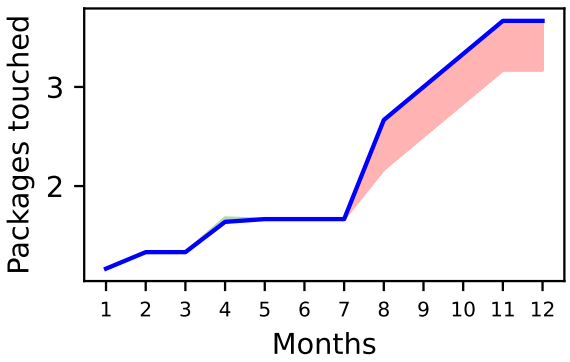


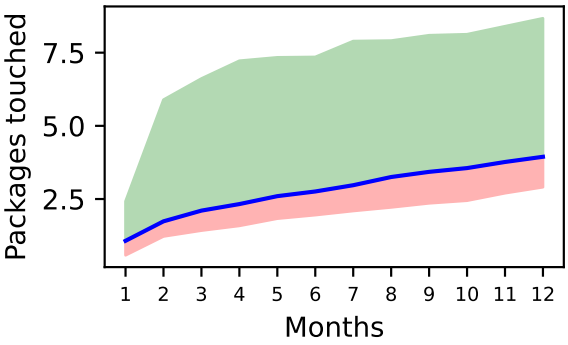
Figure 28: A time series of the number of months (x-axis) against the average (mean) total packages touched (y-axis) for all developers 204,



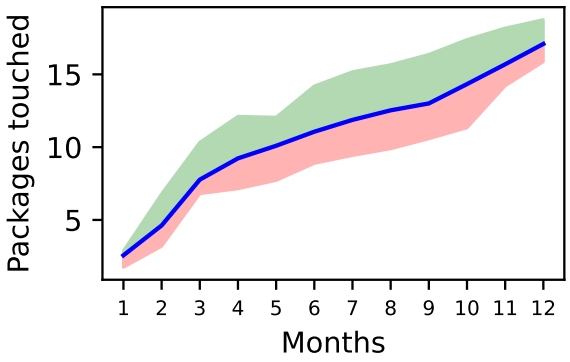
(a) Transient later joiner developers (113) showing packages touched



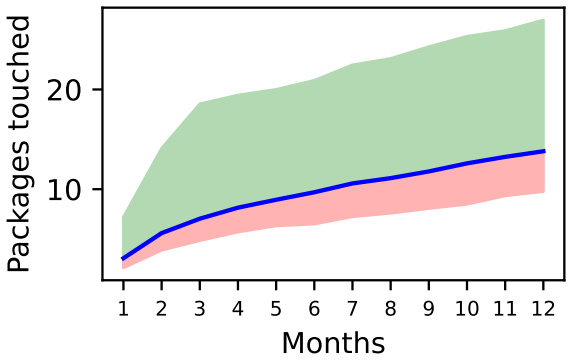
(b) Moderate founder developers (3) showing packages touched



(c) Moderate later joiner developers (93) showing packages touched



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



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

Figure 29: A time series of the number of month (x-axis) against the average (mean) total packages touched (y-axis) for six categories of developer,

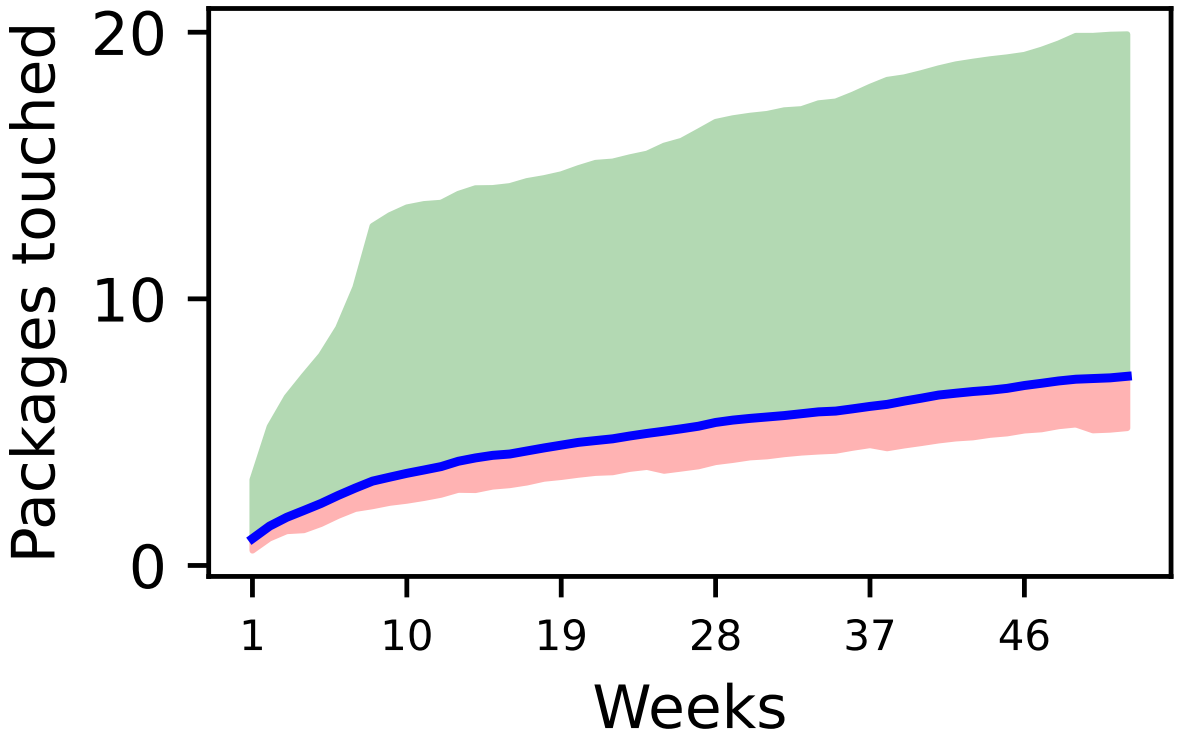
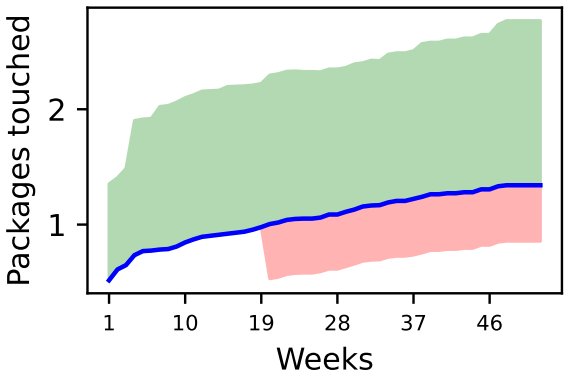
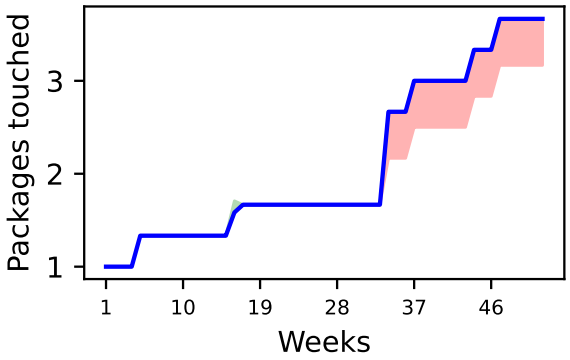


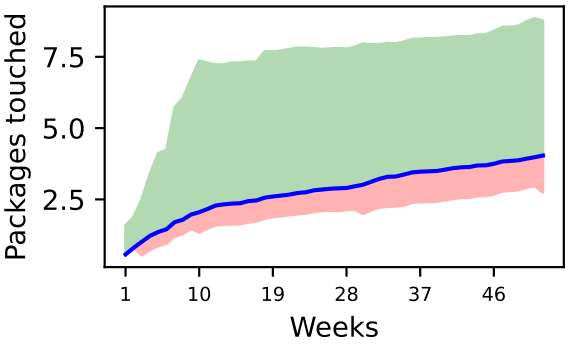
Figure 30: A time series of the number of weeks (x-axis) against the average (mean) total packages touched (y-axis) for all developers 204,



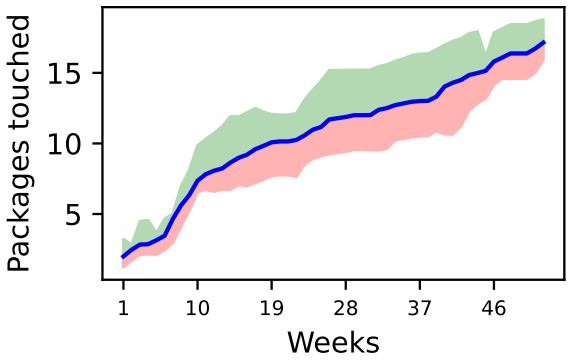
(a) Transient later joiner developers (113) showing packages touched



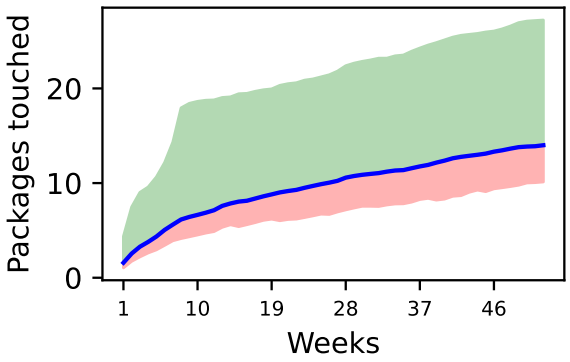
(b) Moderate founder developers (3) showing packages touched



(c) Moderate later joiner developers (93) showing packages touched



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



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

Figure 31: A time series of the number of week (x-axis) against the average (mean) total packages touched (y-axis) for six categories of developer,

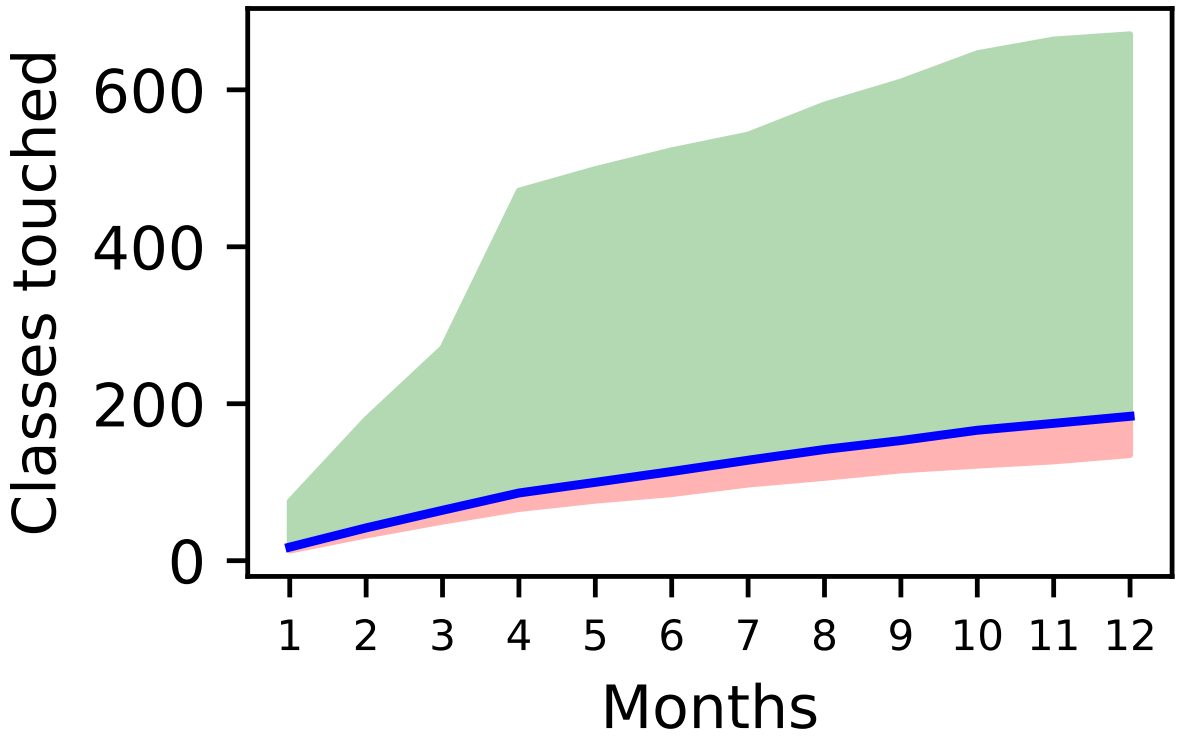
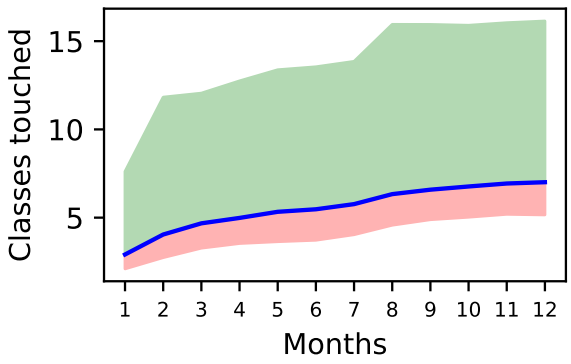
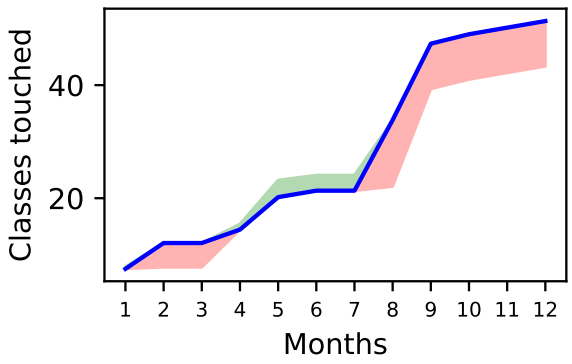


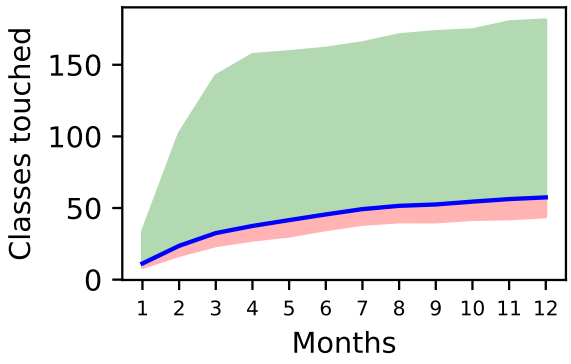
Figure 32: A time series of the number of months (x-axis) against the average (mean) total classes touched (y-axis) for all developers 204,



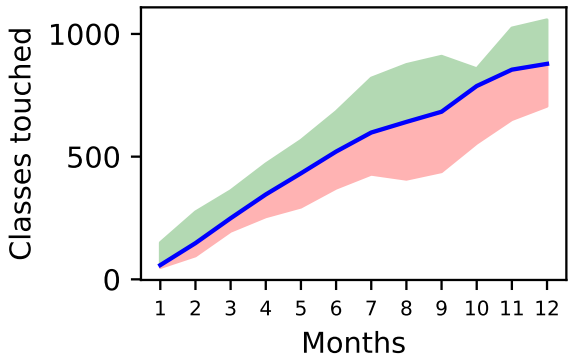
(a) Transient later joiner developers (113) showing classes touched



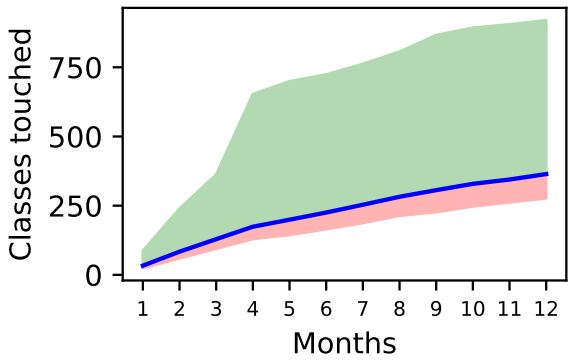
(b) Moderate founder developers (3) showing classes touched



(c) Moderate later joiner developers (93) showing classes touched



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



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

Figure 33: A time series of the number of month (x-axis) against the average (mean) total classes touched (y-axis) for six categories of developer,

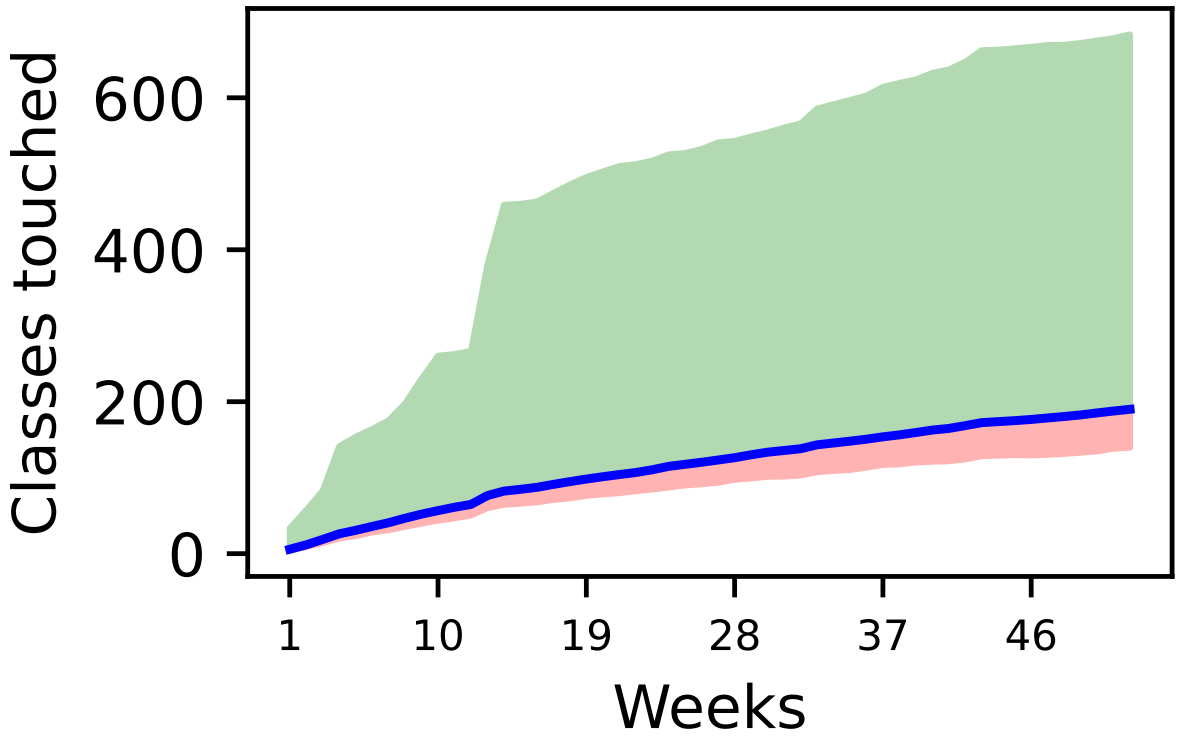
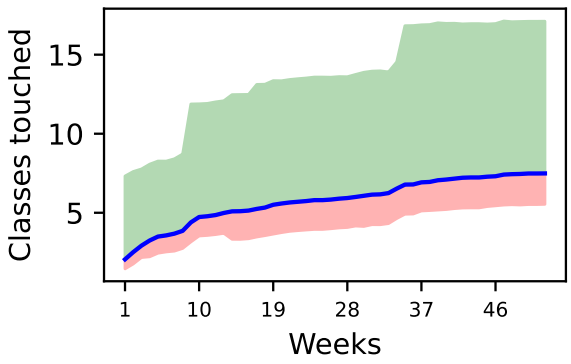
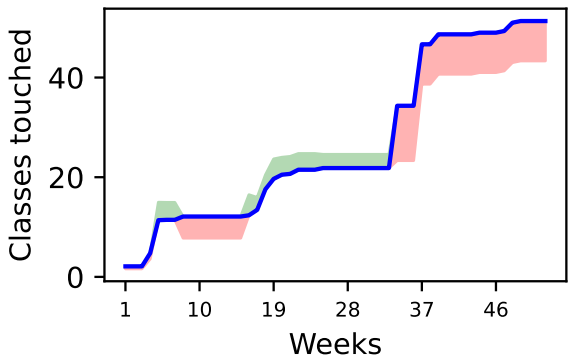


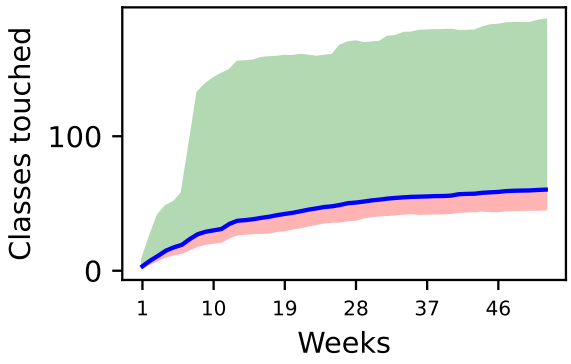
Figure 34: A time series of the number of weeks (x-axis) against the average (mean) total classes touched (y-axis) for all developers 204,



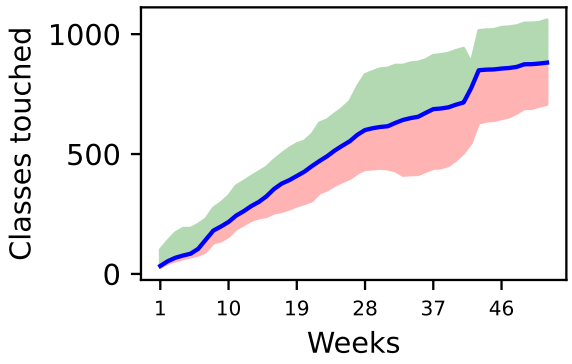
(a) Transient later joiner developers (113) showing classes touched



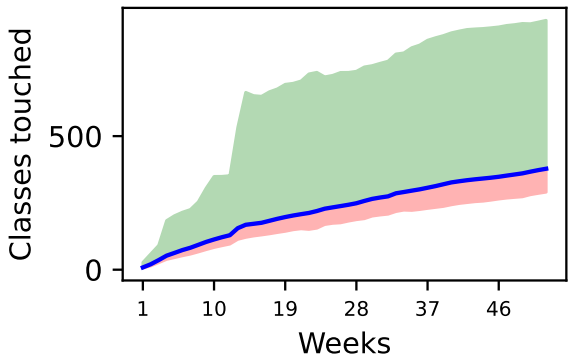
(b) Moderate founder developers (3) showing classes touched



(c) Moderate later joiner developers (93) showing classes touched



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



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

Figure 35: A time series of the number of week (x-axis) against the average (mean) total classes touched (y-axis) for six categories of developer,

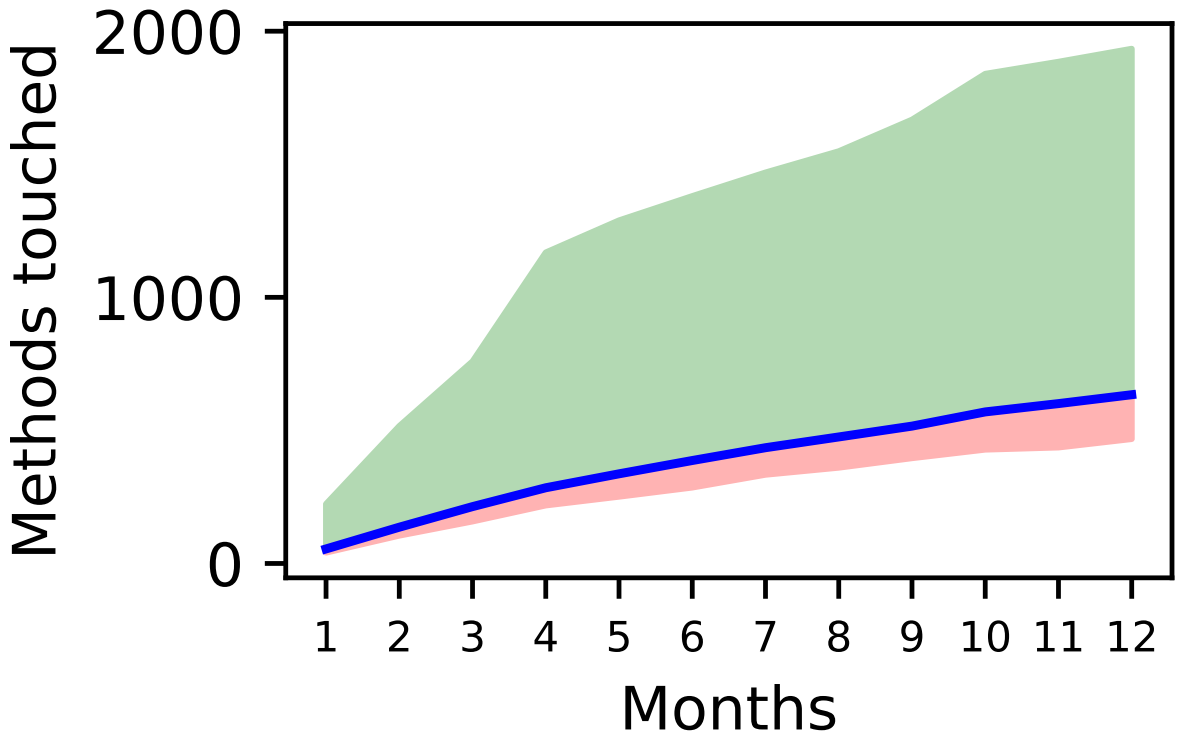
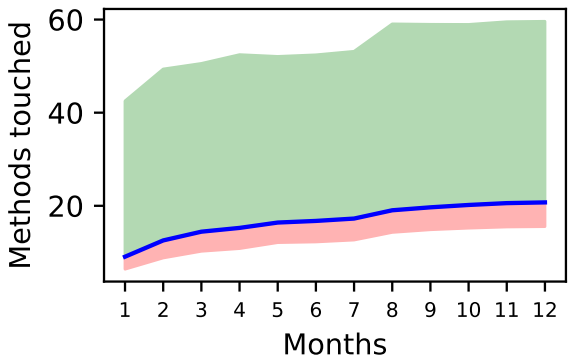
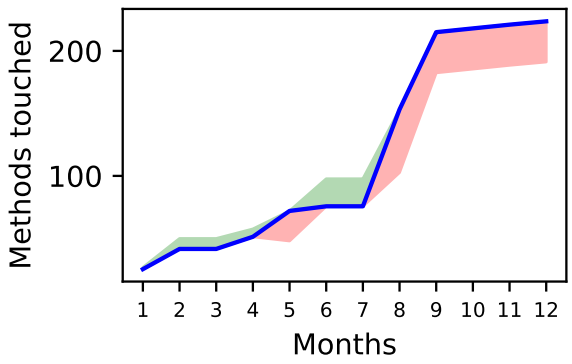


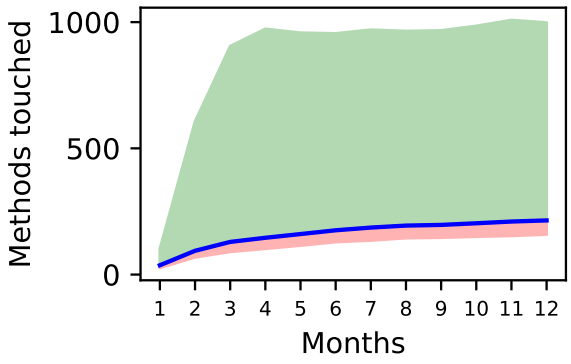
Figure 36: A time series of the number of months (x-axis) against the average (mean) total methods touched (y-axis) for all developers 204,



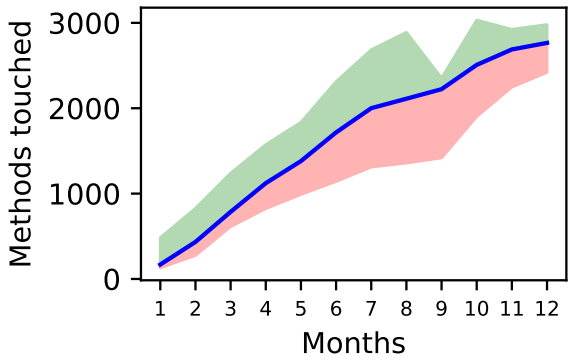
(a) Transient later joiner developers (113) showing methods touched



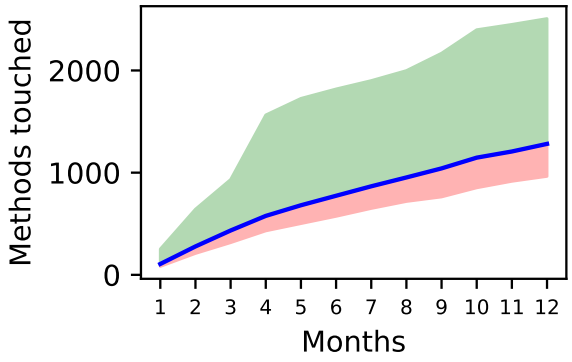
(b) Moderate founder developers (3) showing methods touched



(c) Moderate later joiner developers (93) showing methods touched



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



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

Figure 37: A time series of the number of month (x-axis) against the average (mean) total methods touched (y-axis) for six categories of developer,

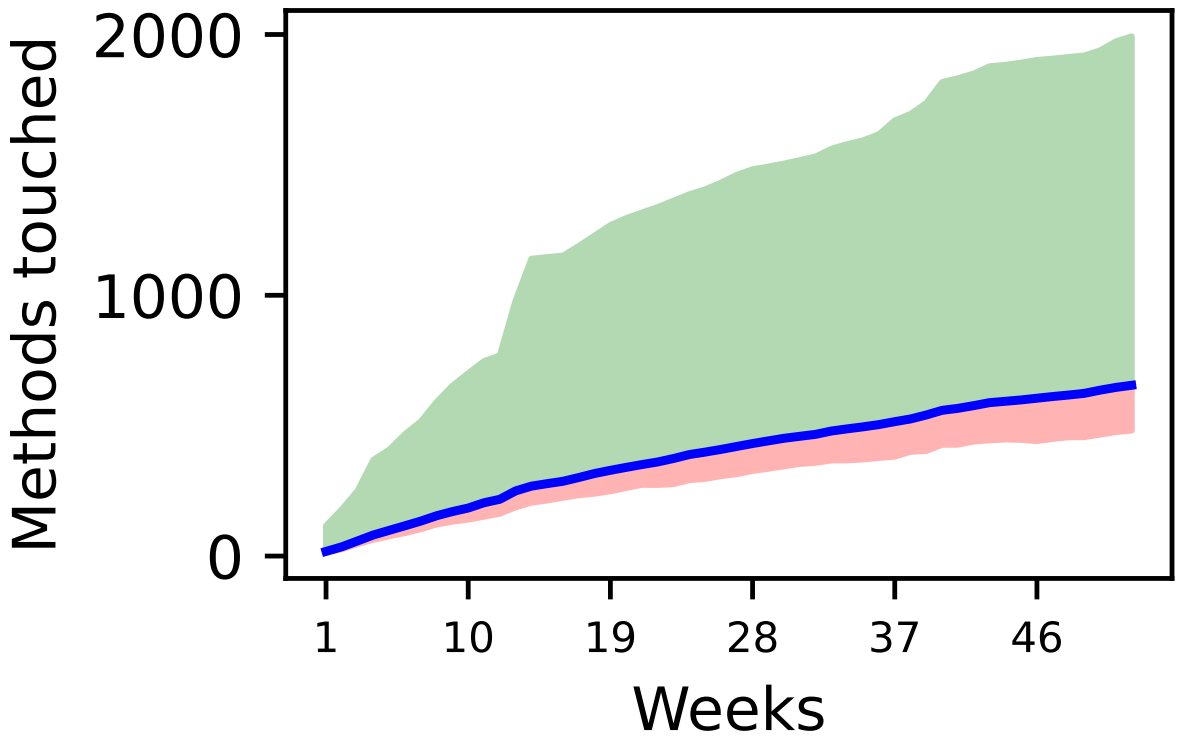
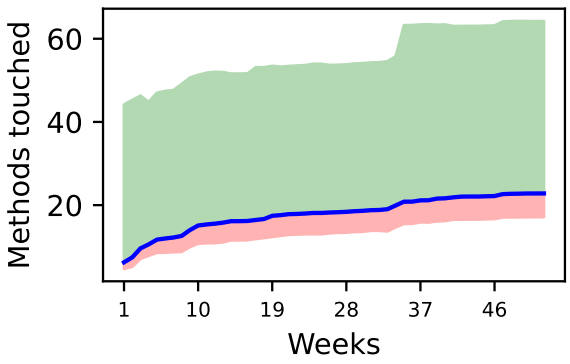
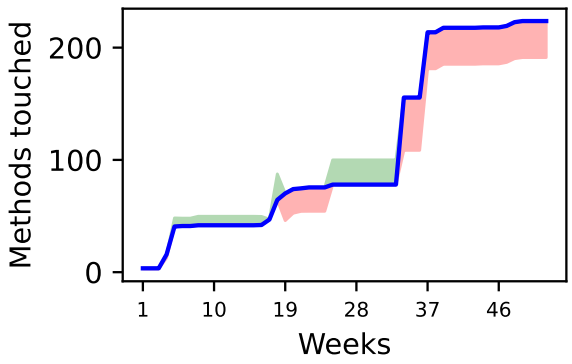


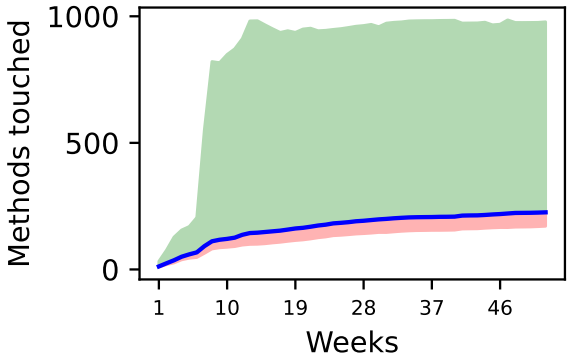
Figure 38: A time series of the number of weeks (x-axis) against the average (mean) total methods touched (y-axis) for all developers 204,



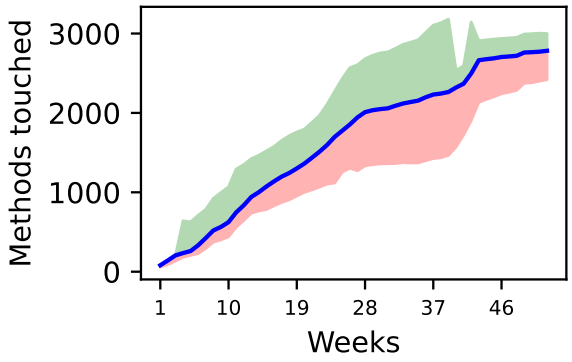
(a) Transient later joiner developers (113) showing methods touched



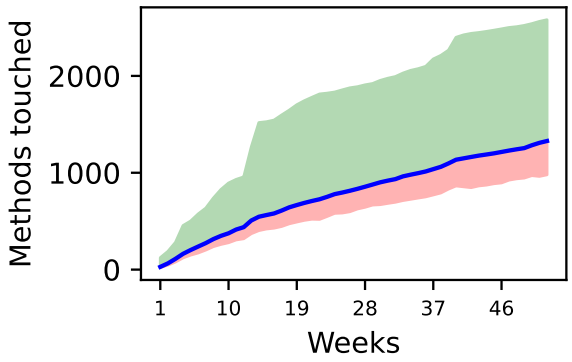
(b) Moderate founder developers (3) showing methods touched



(c) Moderate later joiner developers (93) showing methods touched



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



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

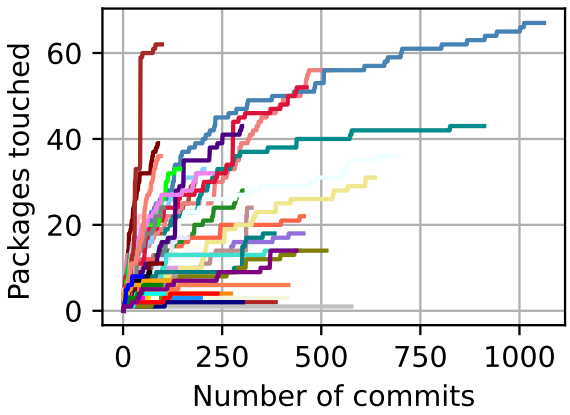
Figure 39: A time series of the number of week (x-axis) against the average (mean) total methods touched (y-axis) for six categories of developer,

3.2 Scatter components touched by developer - 21 For developer commits and components touched.

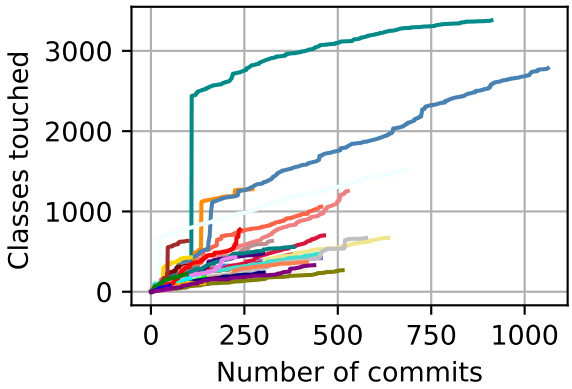
A scatter plot for the first 50 sustained late joiner developers showing the number of commits to components touched.

Table 3: Table of first 50 developers in the Scatter Developer graphs.

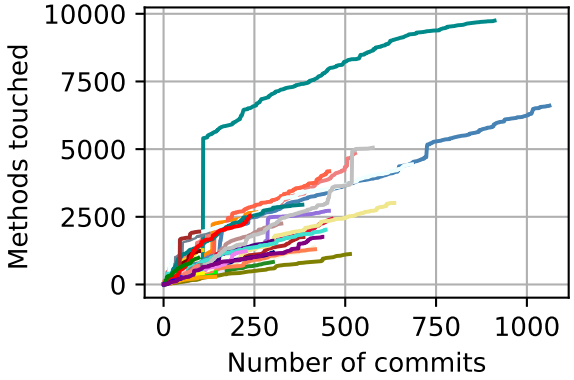
Developer ID	Colour
199	Maroon
6252	Sienna
6253	LightCoral
6259	DarkCyan
6260	DarkOrange
6262	MediumPurple
6263	LightSkyBlue
6264	HotPink
6267	ForestGreen
6271	FireBrick
6272	DodgerBlue
6276	Tomato
6278	SeaGreen
6280	SlateGray
6282	Ivory
6284	Orchid
6288	Chocolate
6295	SteelBlue
6297	RosyBrown
6298	Azure
6302	Tan
6306	Plum
6309	Crimson
6312	Khaki
6313	Lavender
6314	Indigo
6315	Turquoise
6326	Beige
6329	Silver
6333	Gold
6338	Coral
6342	Lime
6343	Maroon
6349	Navy
6352	Teal
6355	Olive
6363	Violet
6368	Gray
6369	Pink
6370	Brown
6373	Black
6374	Yellow
6375	Magenta
6379	Cyan
6391	Salmon
6410	Orange
6411	Red
6417	Green
6418	Blue
6420	Purple



(a) packages



(b) classes

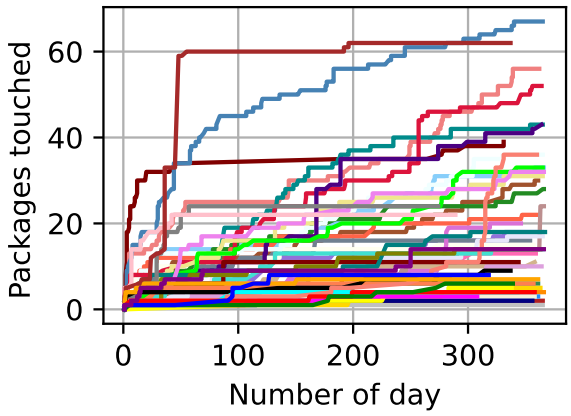


(c) methods

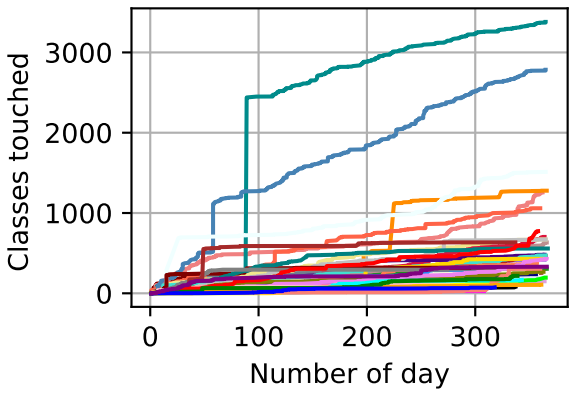
Figure 40: First fifty 50 sustained later joiner developers from this repository. The number of commits against the number of components touched. Developers with colours are shown in Table 3.

3.3 Scatter components touched by developer - 21 For developer day and components touched.

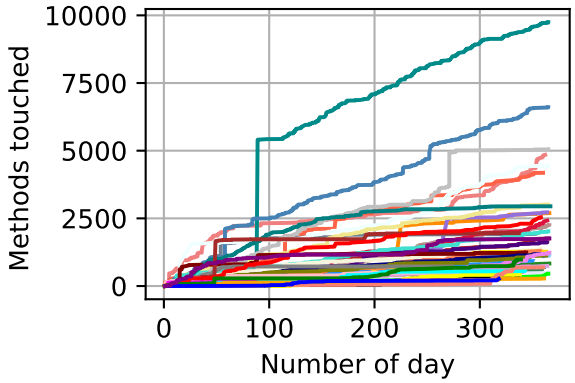
A scatter plot for the first 50 sustained late joiner developers showing the number of day to components touched.



(a) packages



(b) classes



(c) methods

Figure 41: First fifty 50 sustained later joiner developers from this repository. The number of day against the number of components touched. Developers with colours are shown in Table 3.