Learning a Codebase

Derek Somerville

December 8, 2025

Contents

1 Repository

1.1 Glossary - Summary

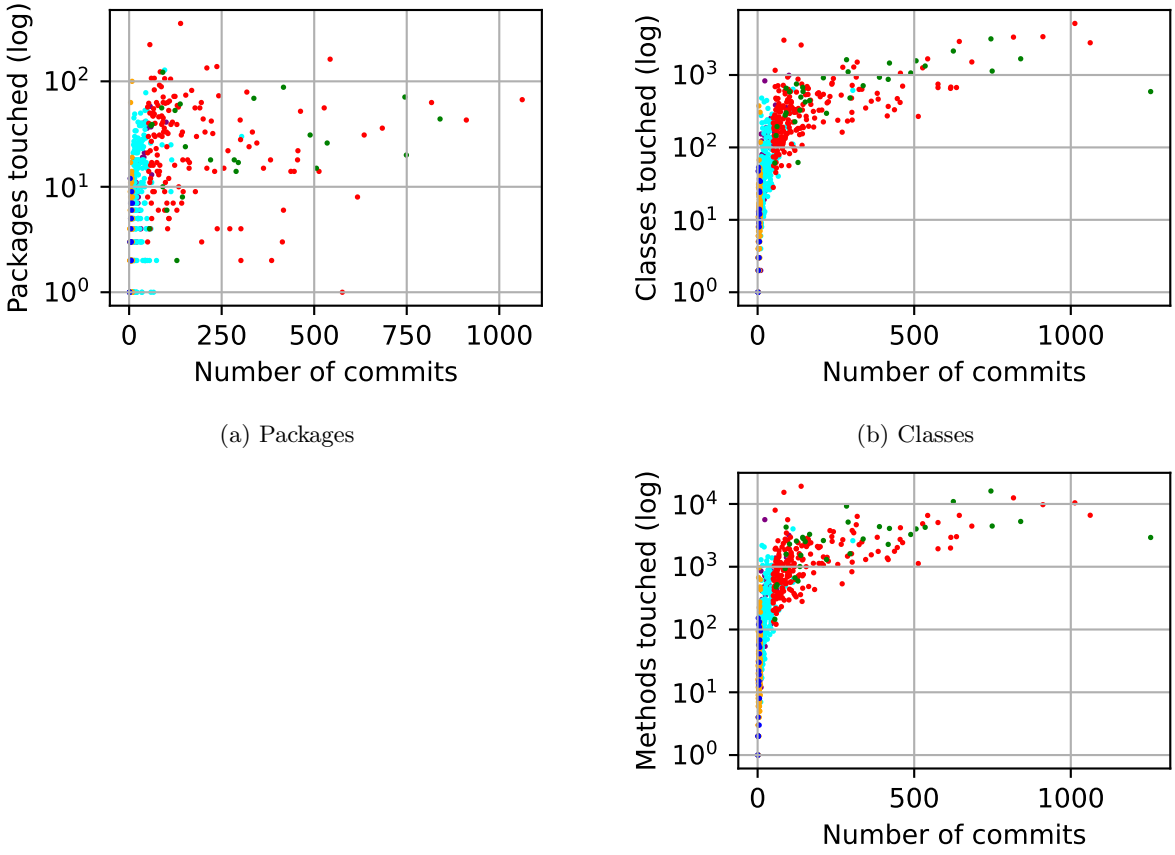
- The founder developer starts in the first six months of a project.
- Late joiner developers begin after six months.
- Transient developers have ten (10) or fewer commits.
- Moderate developers have more than ten (10) commits but fewer than 50 commits or commits for less than 250 days.
- Sustained developers 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 (6) 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

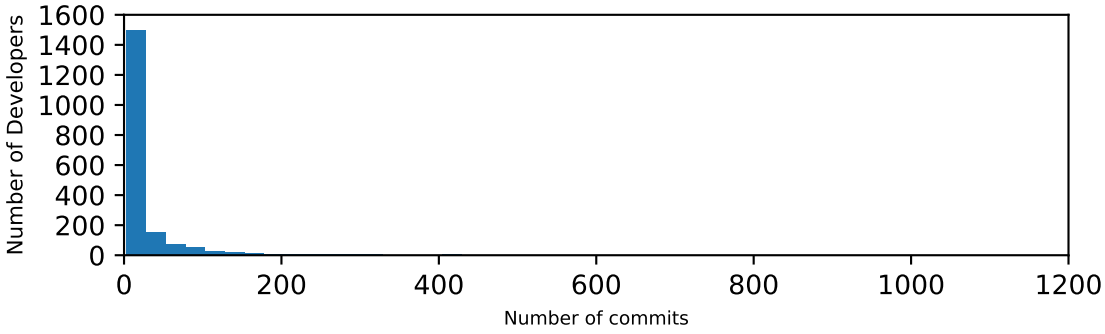
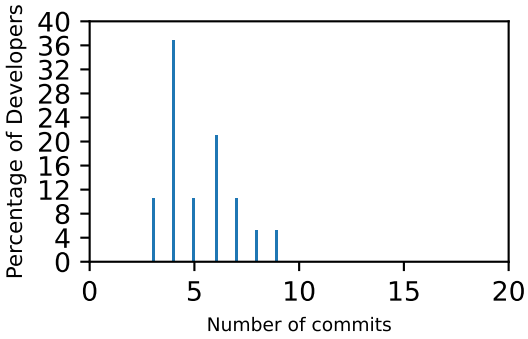
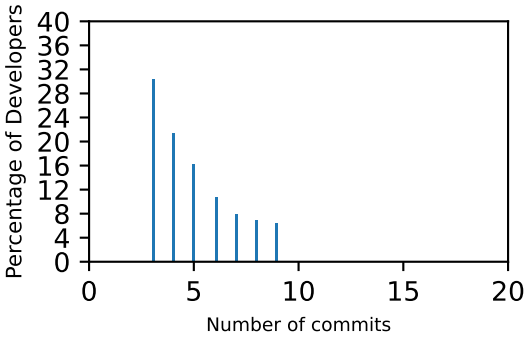


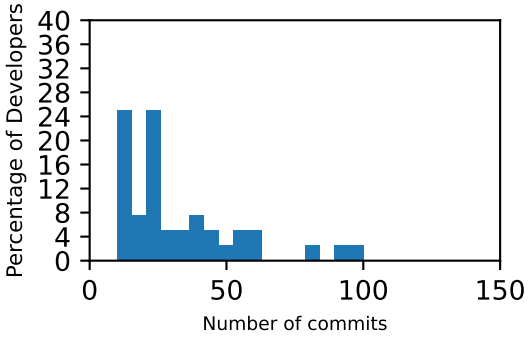
Figure 2: Histogram of the number from all developers (n=1906) (y-axis) against the total number of commits (x-axis) in 15 repositories sampled from GitHub, excluding developers with less than three (3) commits.



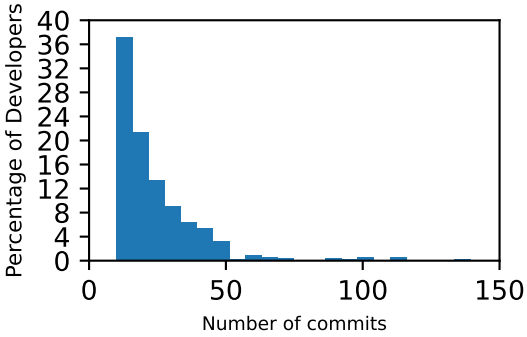
(a) Transient founder (n=19).



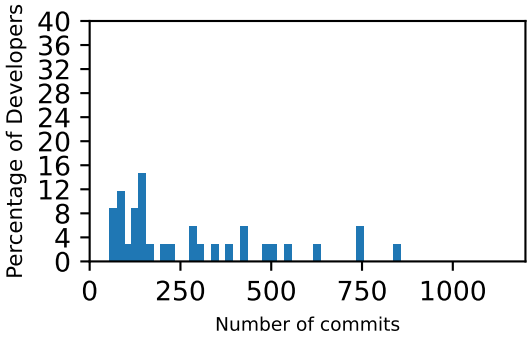
(b) Transient later joiner (n=1030).



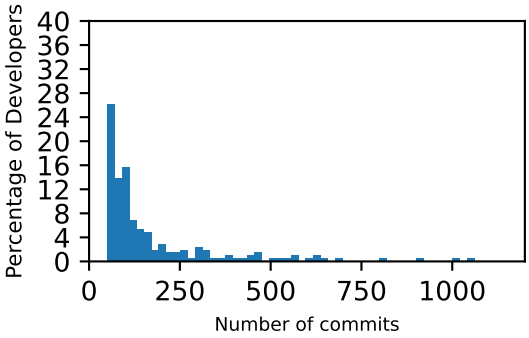
(c) Moderate founder (n=40).



(d) Moderate later joiner (n=580).



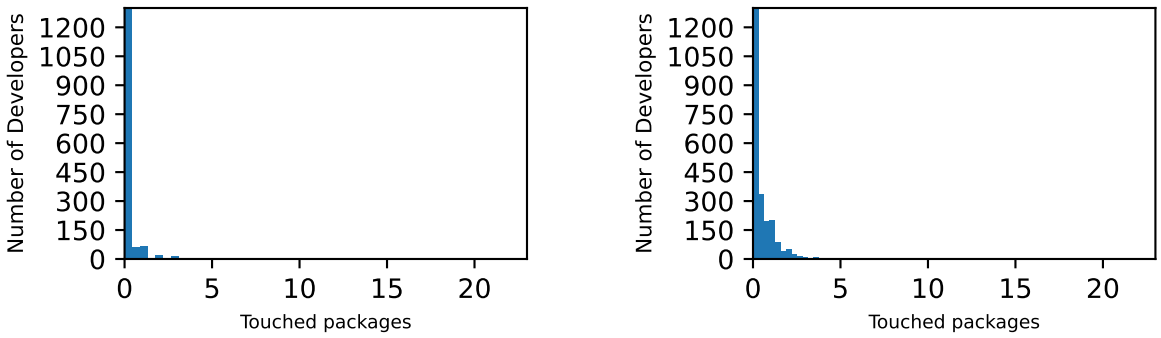
(e) Sustained founder (n=34).



(f) Sustained later joiner (n=203).

Figure 3: Histogram of percentage of developers (y-axis) against total commits made (x-axis) in six (6) categories. This is from 15 repositories sampled from GitHub, excluding developers with less than three (3) commits.

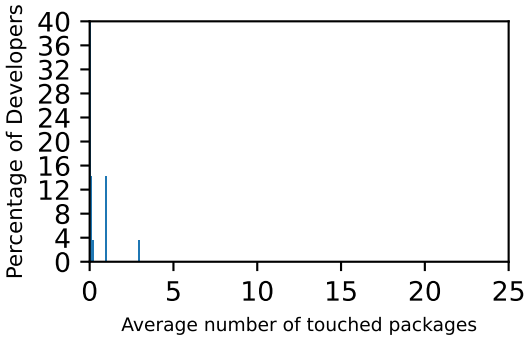
1.5 Histogram components touched by developer - For components with total



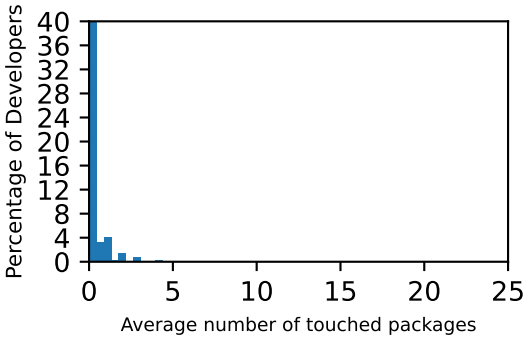
(a) By Day.

(b) By Commit.

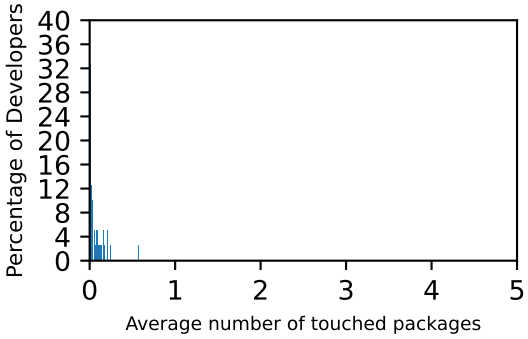
Figure 4: Histogram of the number of developers (y-axis) against the average packages touched by day and commit (x-axis) from all 2278 developers from 15 repositories sampled from GitHub.



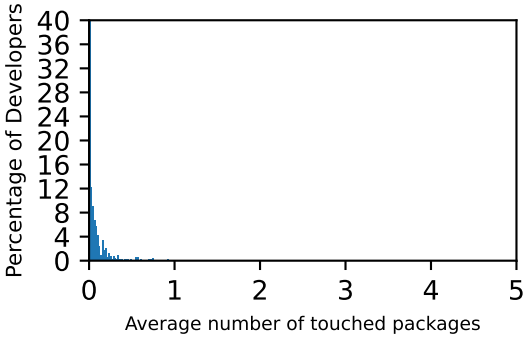
(a) Transient founder (n=28).



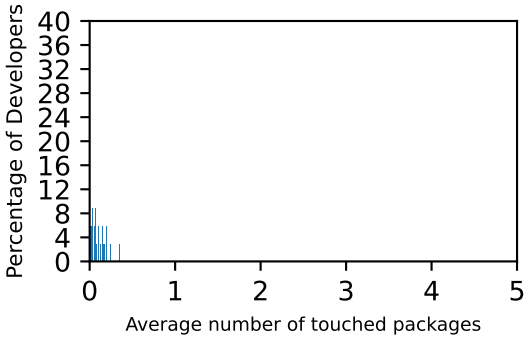
(b) Transient later joiner (n=1396).



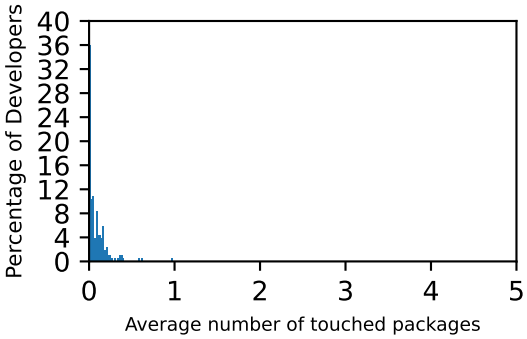
(c) Moderate founder (n=40).



(d) Moderate later joiner (n=580).

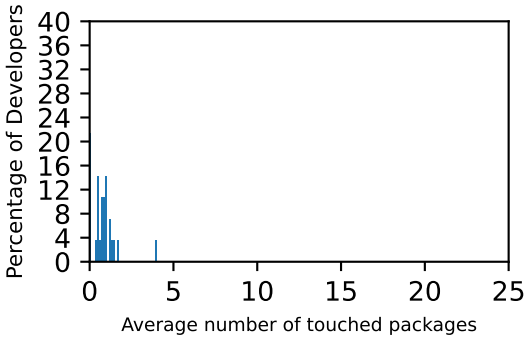


(e) Sustained founder (n=34).

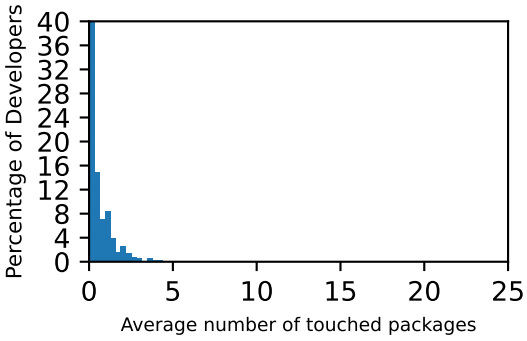


(f) Sustained later joiner (n=203).

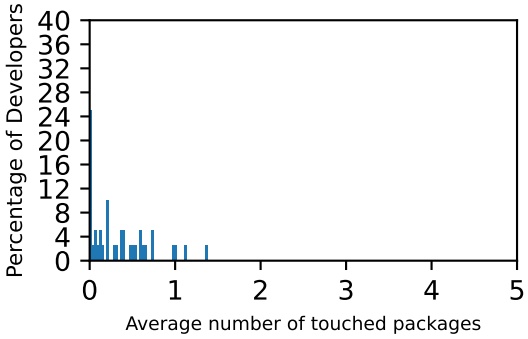
Figure 5: Histogram of the number of developers (y-axis) against the average packages touched by day (x-axis) from six (6) categories of developers from 15 repositories sampled from GitHub.



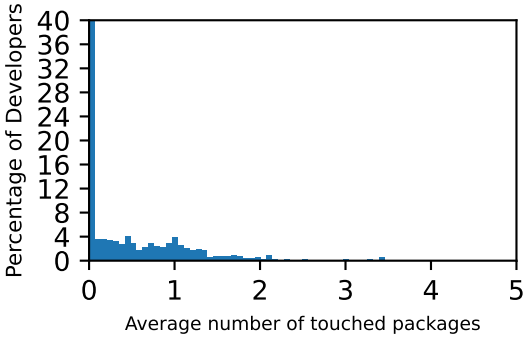
(a) Transient founder (n=28).



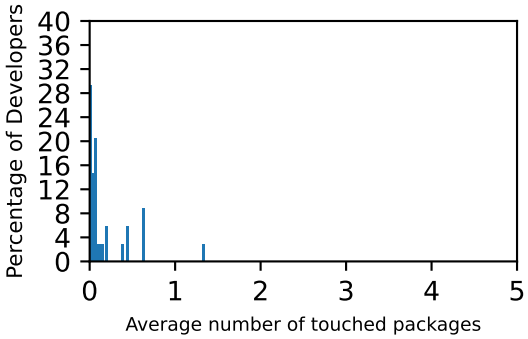
(b) Transient later joiner (n=1396).



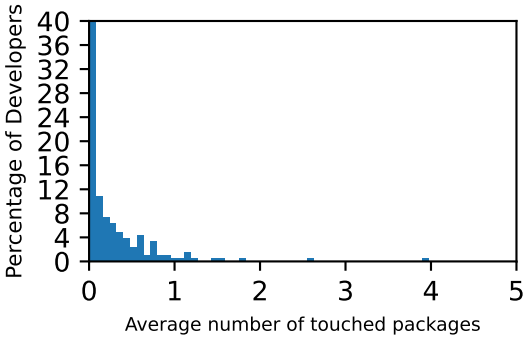
(c) Moderate founder (n=40).



(d) Moderate later joiner (n=580).

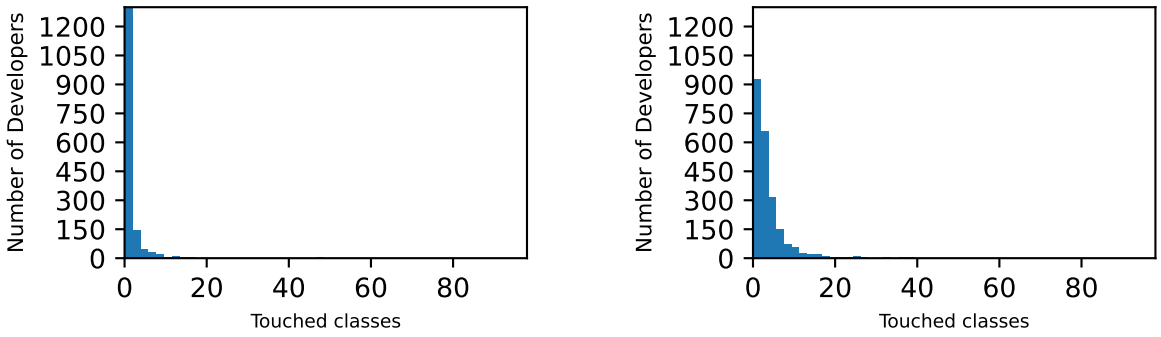


(e) Sustained founder (n=34).



(f) Sustained later joiner (n=203).

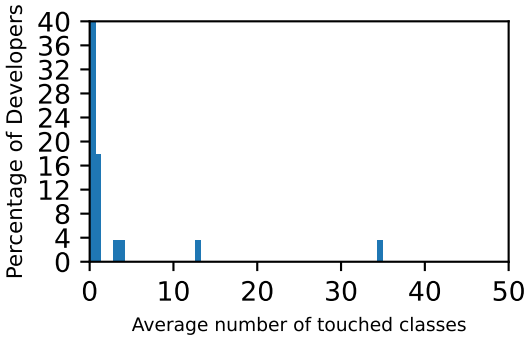
Figure 6: Histogram of the number of developers (y-axis) against the average packages touched by commit (x-axis) from six (6) categories of developers from 15 repositories sampled from GitHub.



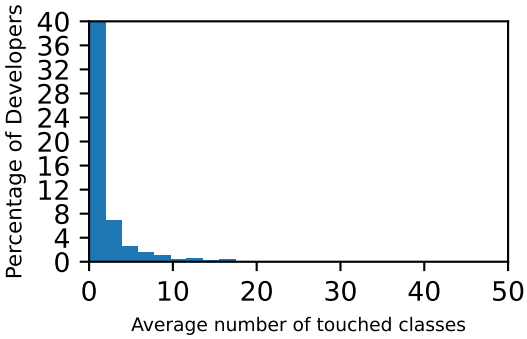
(a) By Day.

(b) By Commit.

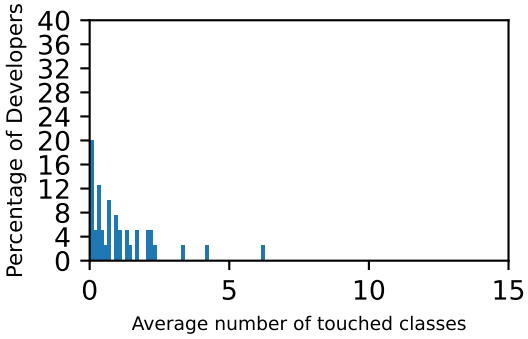
Figure 7: Histogram of the number of developers (y-axis) against the average classes touched by day and commit (x-axis) from all 2278 developers from 15 repositories sampled from GitHub.



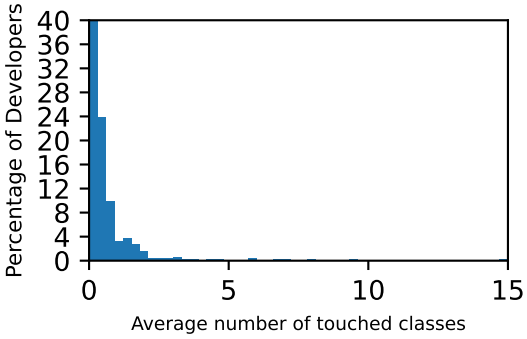
(a) Transient founder (n=28).



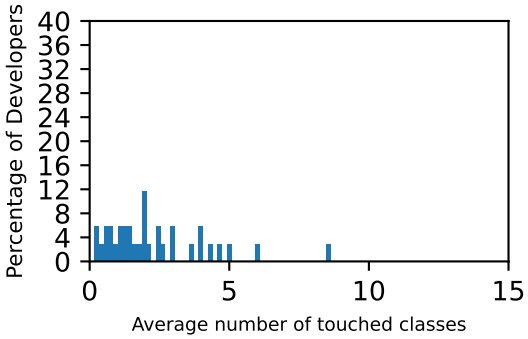
(b) Transient later joiner (n=1396).



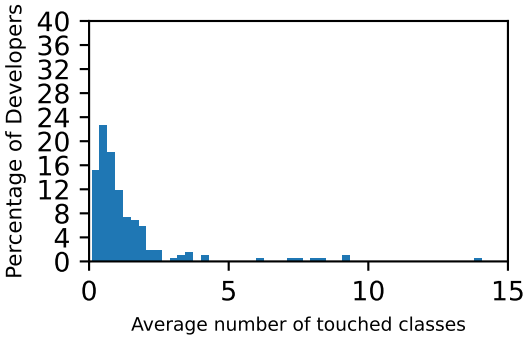
(c) Moderate founder (n=40).



(d) Moderate later joiner (n=580).

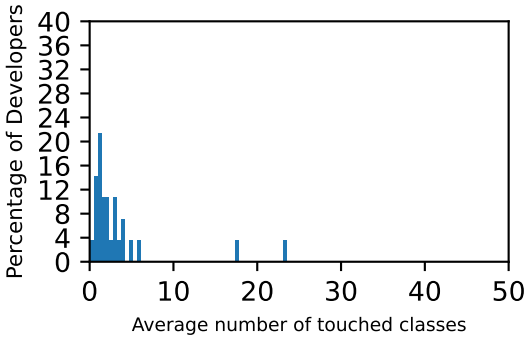


(e) Sustained founder (n=34).

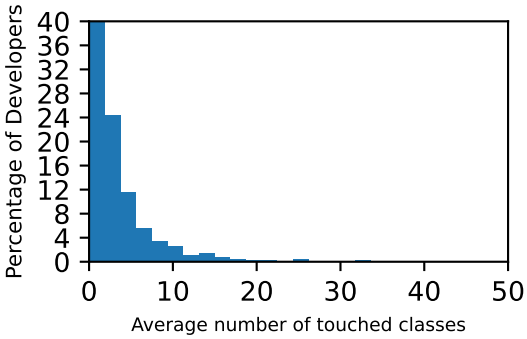


(f) Sustained later joiner (n=203).

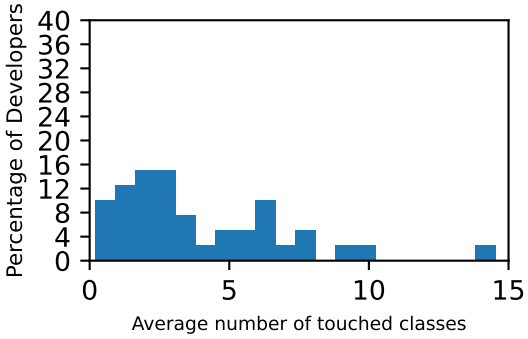
Figure 8: Histogram of the number of developers (y-axis) against the average classes touched by day (x-axis) from six (6) categories of developers from 15 repositories sampled from GitHub.



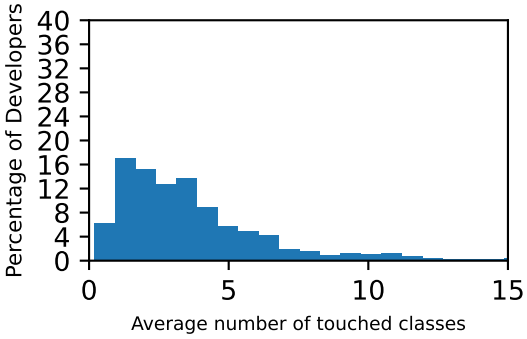
(a) Transient founder (n=28).



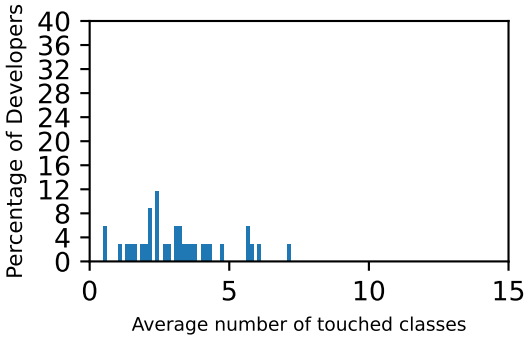
(b) Transient later joiner (n=1396).



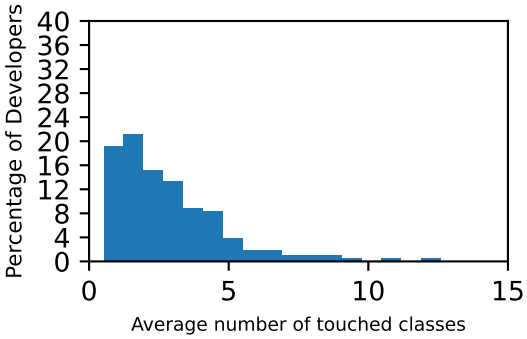
(c) Moderate founder (n=40).



(d) Moderate later joiner (n=580).

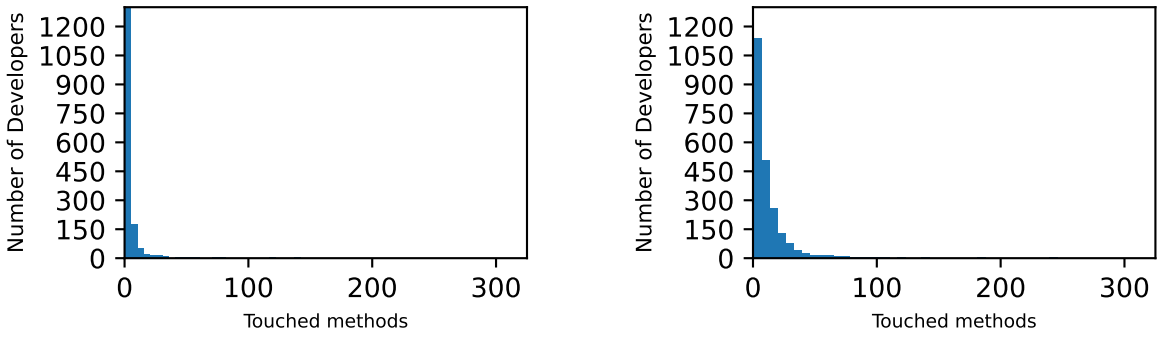


(e) Sustained founder (n=34).



(f) Sustained later joiner (n=203).

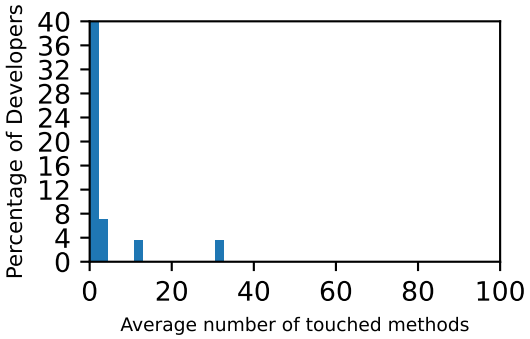
Figure 9: Histogram of the number of developers (y-axis) against the average classes touched by commit (x-axis) from six (6) categories of developers from 15 repositories sampled from GitHub.



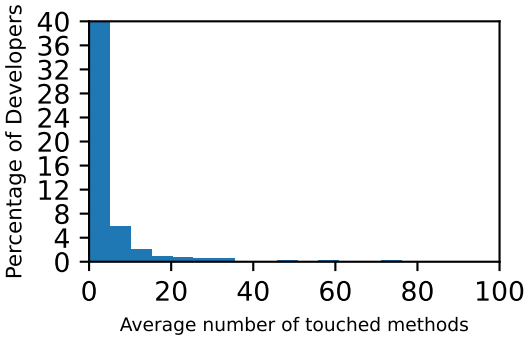
(a) By Day.

(b) By Commit.

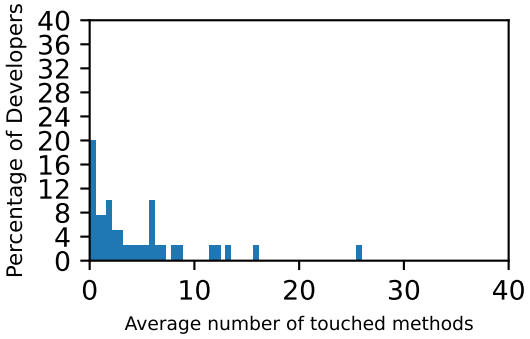
Figure 10: Histogram of the number of developers (y-axis) against the average methods touched by day and commit (x-axis) from all 2278 developers from 15 repositories sampled from GitHub.



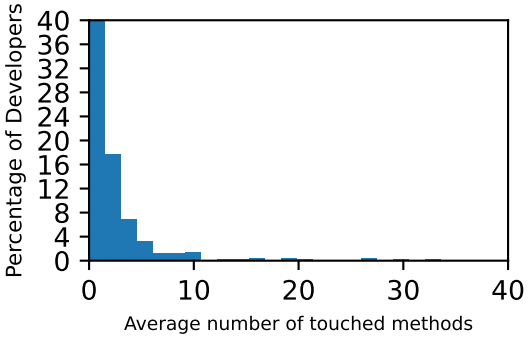
(a) Transient founder (n=28).



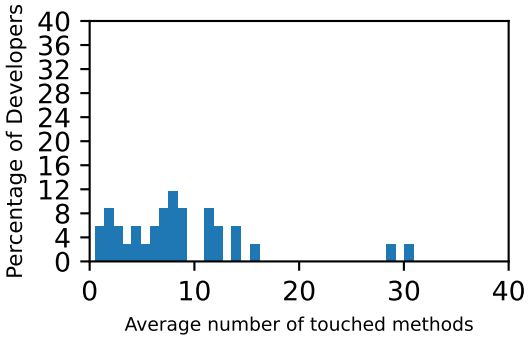
(b) Transient later joiner (n=1396).



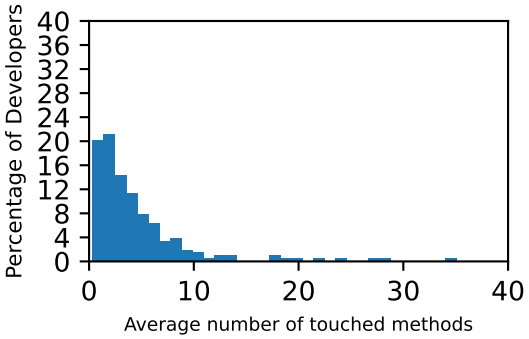
(c) Moderate founder (n=40).



(d) Moderate later joiner (n=580).

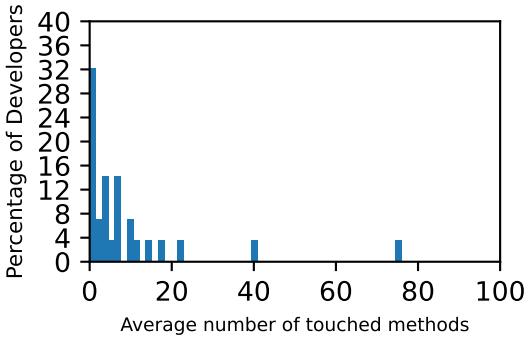


(e) Sustained founder (n=34).

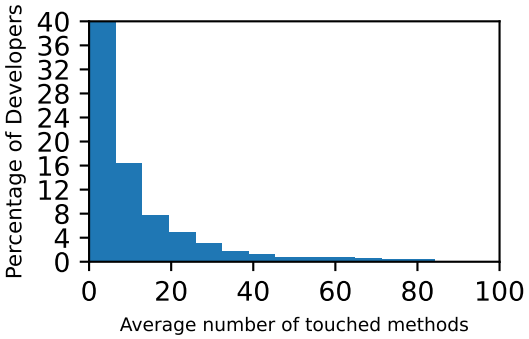


(f) Sustained later joiner (n=203).

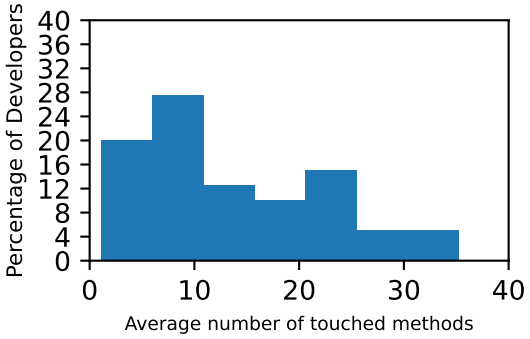
Figure 11: Histogram of the number of developers (y-axis) against the average methods touched by day (x-axis) from six (6) categories of developers from 15 repositories sampled from GitHub.



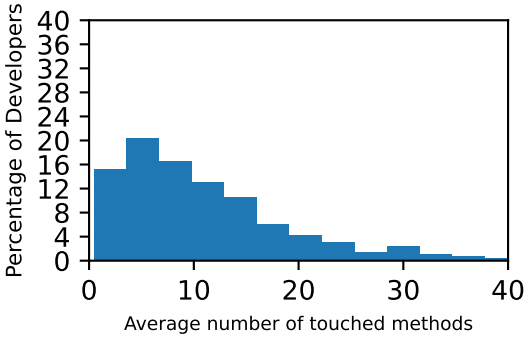
(a) Transient founder (n=28).



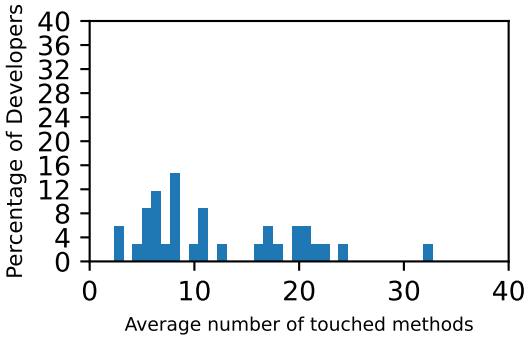
(b) Transient later joiner (n=1396).



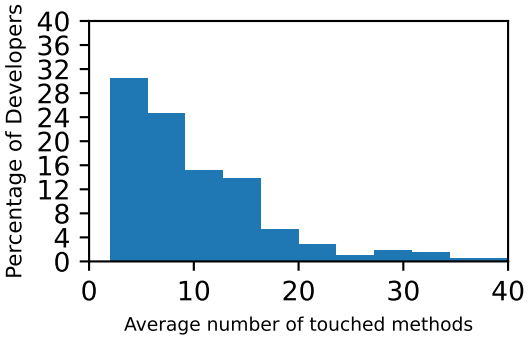
(c) Moderate founder (n=40).



(d) Moderate later joiner (n=580).



(e) Sustained founder (n=34).



(f) Sustained later joiner (n=203).

Figure 12: Histogram of the number of developers (y-axis) against the average methods touched by commit (x-axis) from six (6) categories of developers from 15 repositories sampled from GitHub.

1.6 Time series components touched - For packages touched for each period

A time series of packages touched on average each month.

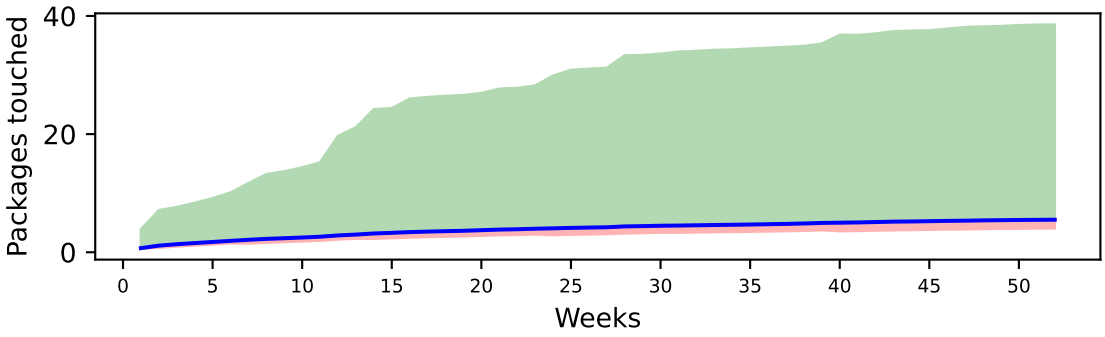
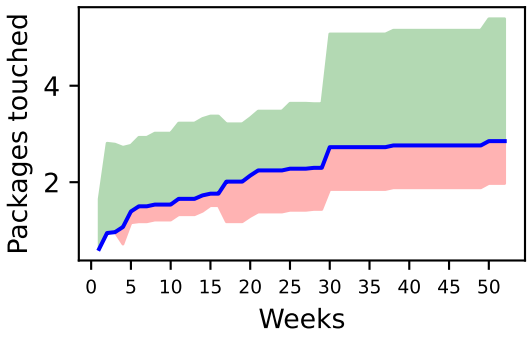
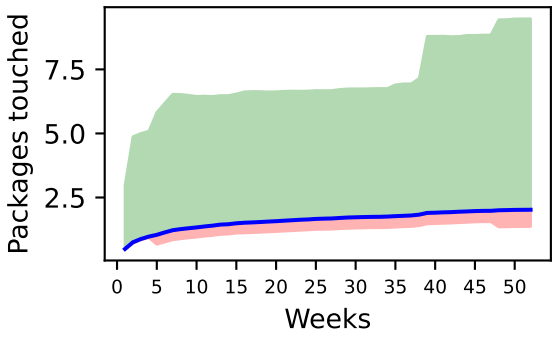


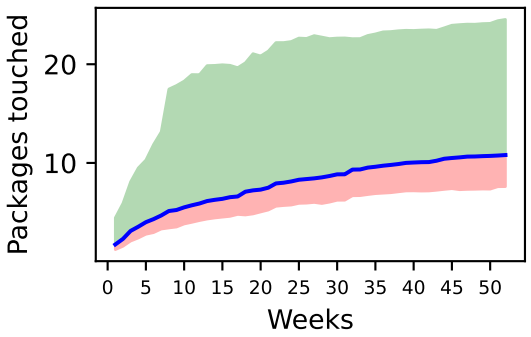
Figure 13: Repository: 0. A time series of the average (mean) total packages touched (y-axis) against the number of weeks (x-axis) for all developers (n=1660).



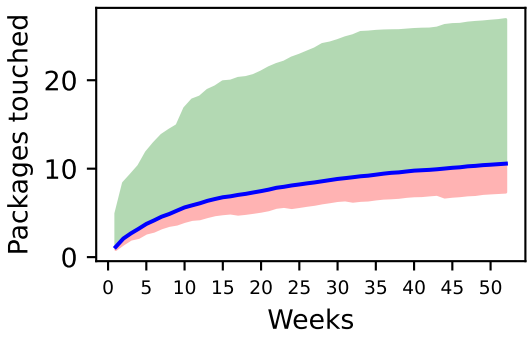
(a) Transient founder developers (n=28)



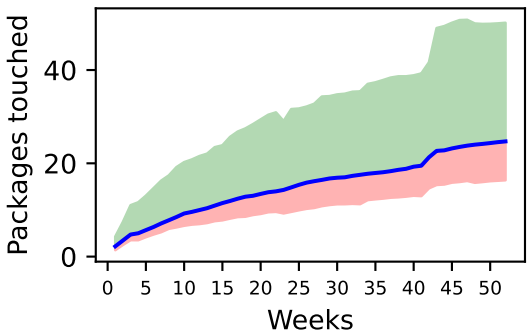
(b) Transient later joiner developers (n=1396)



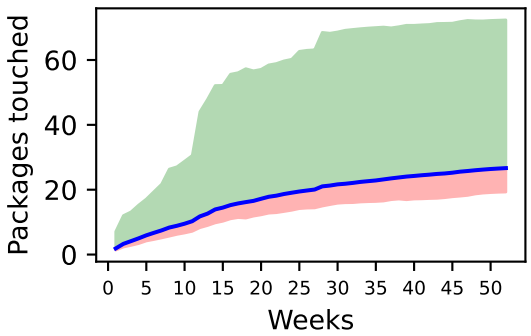
(c) Moderate founder developers (n=40)



(d) Moderate later joiner developers (n=580)



(e) Sustained founder developers (n=34)



(f) Sustained later joiner developers (n=203)

Figure 14: Repository: 0. A time series of the average (mean) total packages touched (y-axis) against the number of week (x-axis) for six (6) categories of developer.

1.7 Time series components touched - For classes touched for each period

A time series of classes touched on average each month.

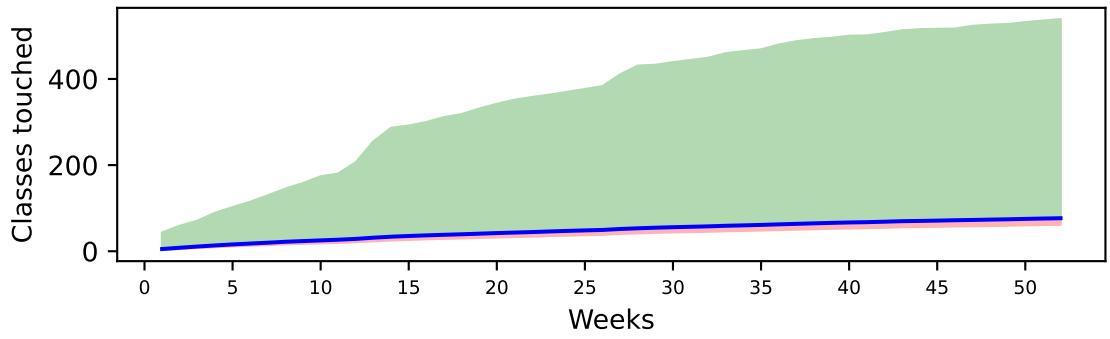
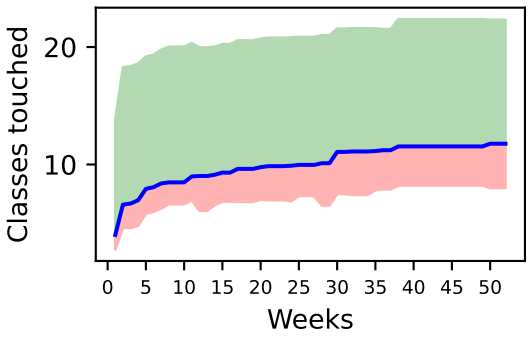
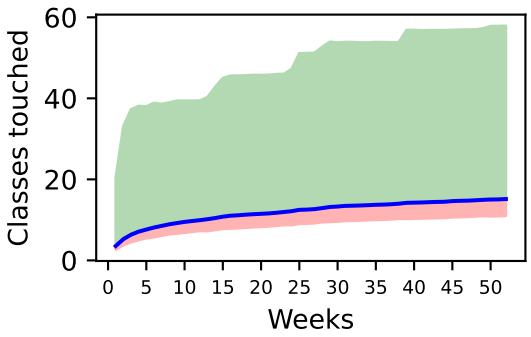


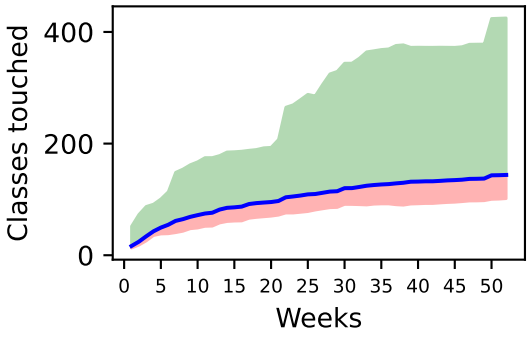
Figure 15: Repository: 0. A time series of the average (mean) total classes touched (y-axis) against the number of weeks (x-axis) for all developers (n=1660).



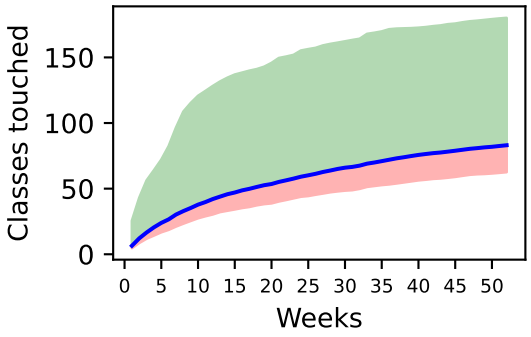
(a) Transient founder developers (n=28)



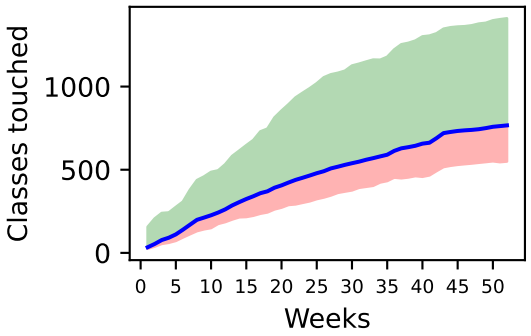
(b) Transient later joiner developers (n=1396)



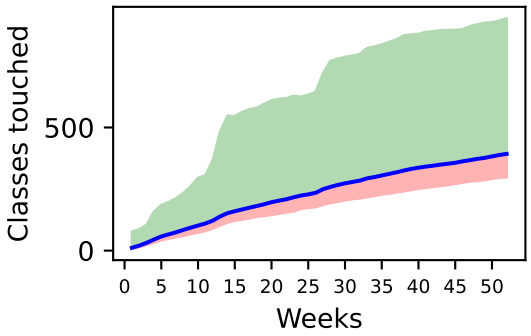
(c) Moderate founder developers (n=40)



(d) Moderate later joiner developers (n=580)



(e) Sustained founder developers (n=34)



(f) Sustained later joiner developers (n=203)

Figure 16: Repository: 0. A time series of the average (mean) total classes touched (y-axis) against the number of week (x-axis) for six (6) categories of developer.

1.8 Time series components touched - For methods touched for each period

A time series of methods touched on average each month.

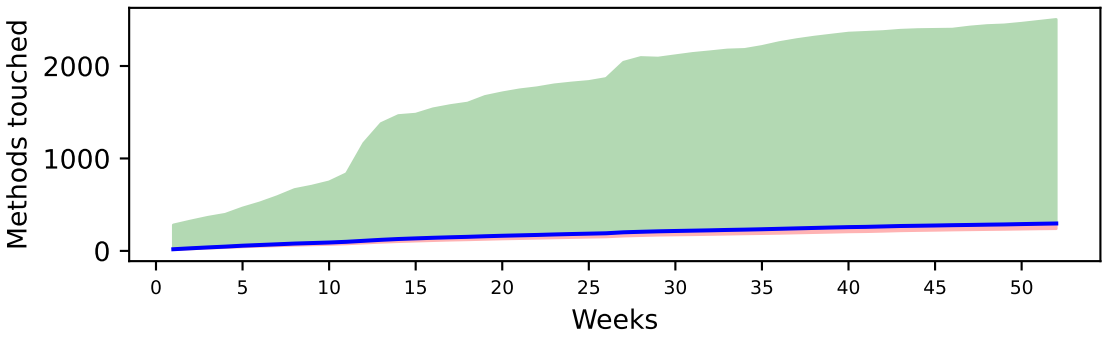
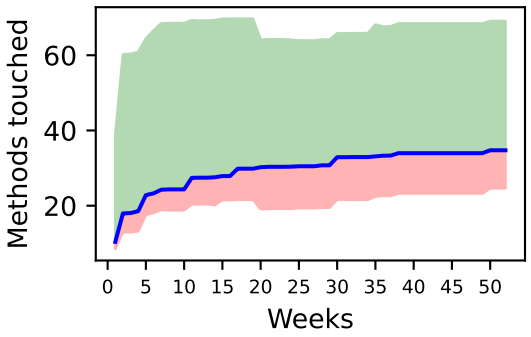
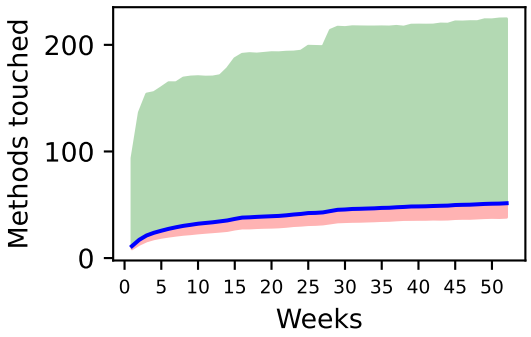


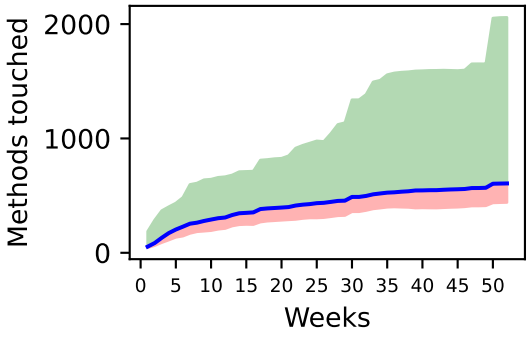
Figure 17: Repository: 0. A time series of the average (mean) total methods touched (y-axis) against the number of weeks (x-axis) for all developers (n=1660).



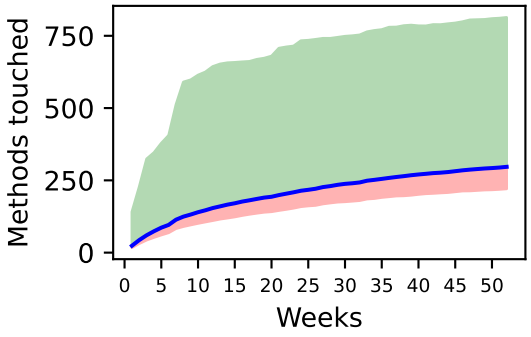
(a) Transient founder developers (n=28)



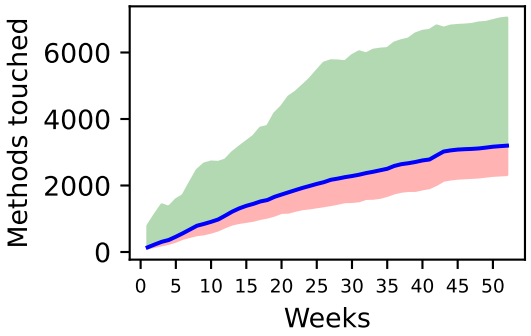
(b) Transient later joiner developers (n=1396)



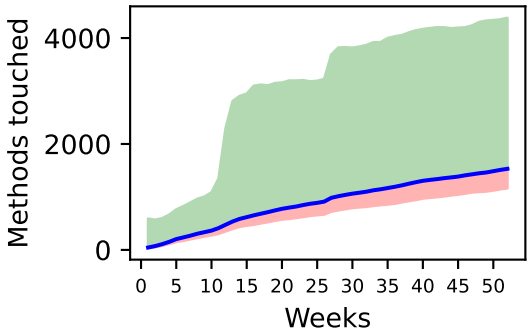
(c) Moderate founder developers (n=40)



(d) Moderate later joiner developers (n=580)



(e) Sustained founder developers (n=34)



(f) Sustained later joiner developers (n=203)

Figure 18: Repository: 0. A time series of the average (mean) total methods touched (y-axis) against the number of week (x-axis) for six (6) categories of developer.

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.

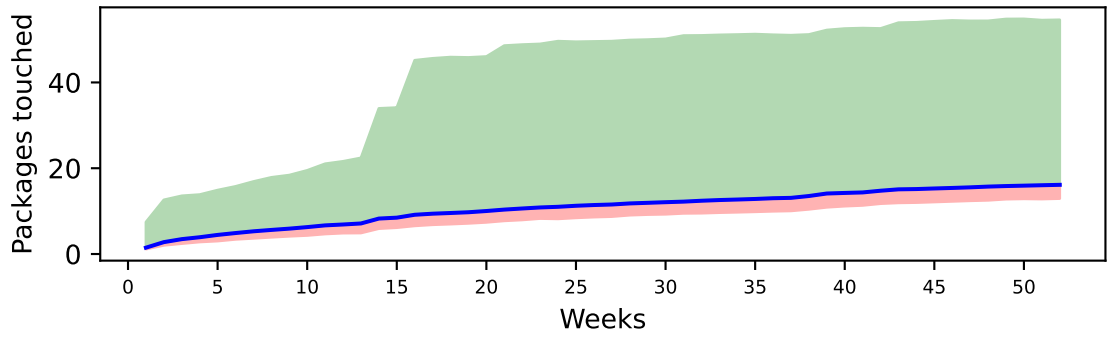
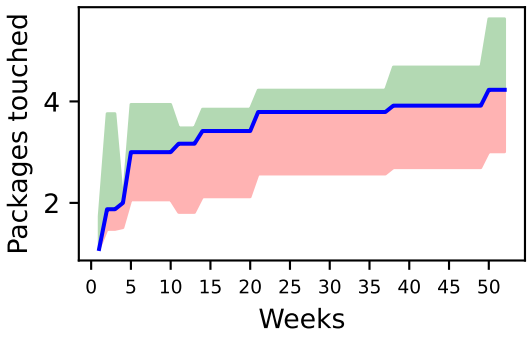
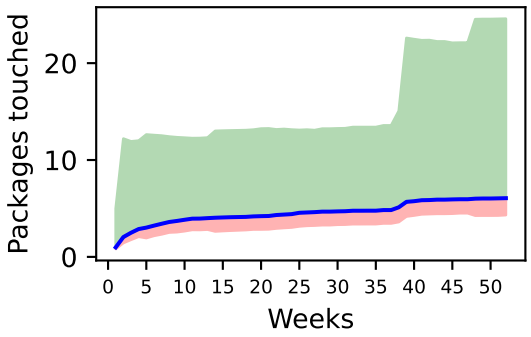


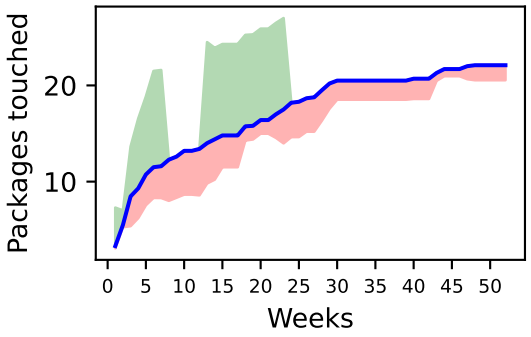
Figure 19: Repository: 14. A time series of the average (mean) total packages touched (y-axis) against the number of weeks (x-axis) for all developers (n=154).



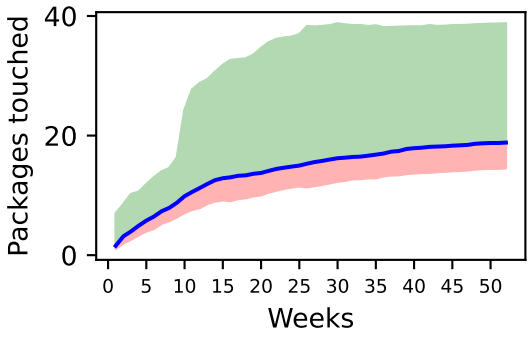
(a) Transient founder developers (n=8)



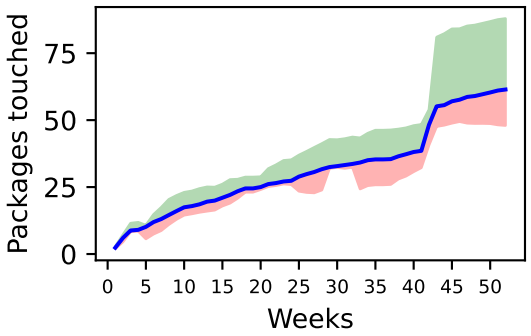
(b) Transient later joiner developers (n=119)



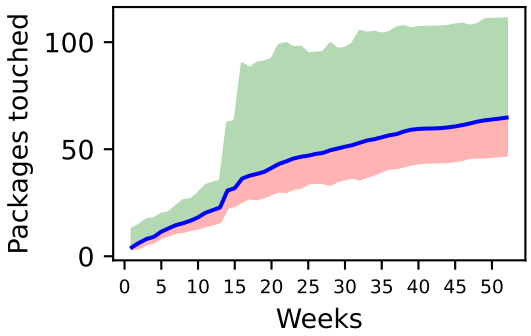
(c) Moderate founder developers (n=5)



(d) Moderate later joiner developers (n=85)



(e) Sustained founder developers (n=6)



(f) Sustained later joiner developers (n=21)

Figure 20: Repository: 14. A time series of the average (mean) total packages touched (y-axis) against the number of week (x-axis) for six (6) categories of developer.

2.2 Time series components touched - 14 For classes touched for each period

A time series of classes touched on average each month.

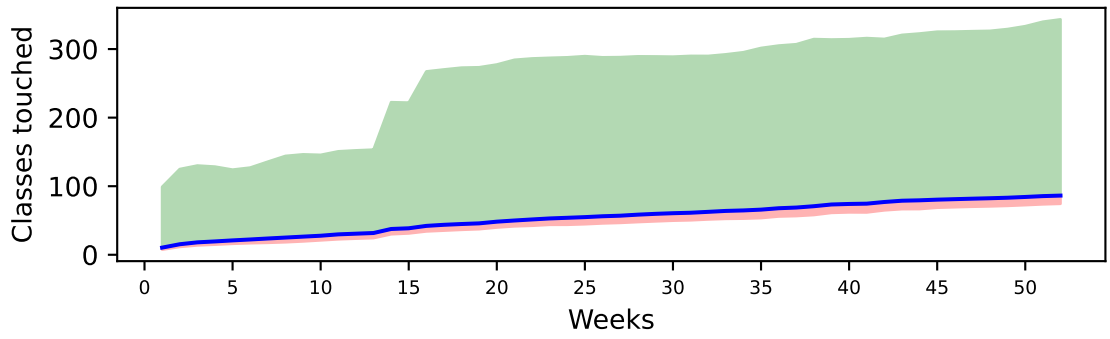
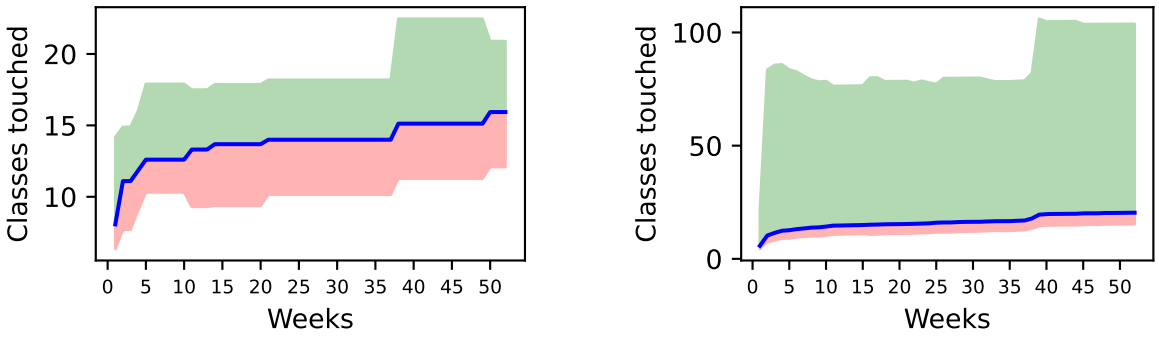
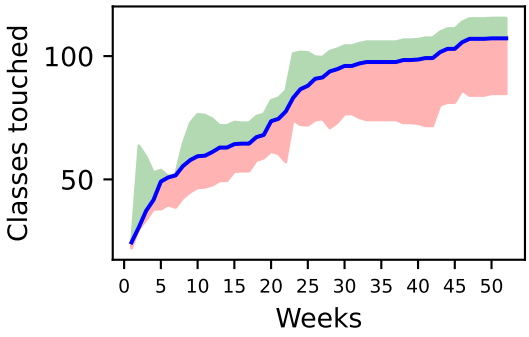


Figure 21: Repository: 14. A time series of the average (mean) total classes touched (y-axis) against the number of weeks (x-axis) for all developers (n=154).

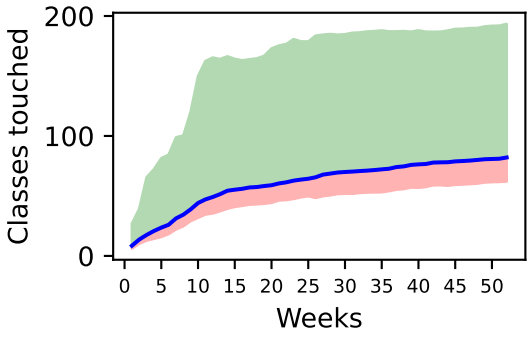


(a) Transient founder developers (n=8)

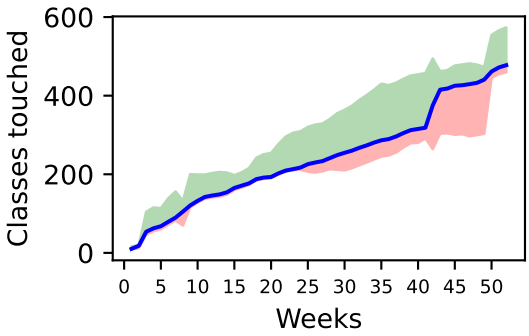
(b) Transient later joiner developers (n=119)



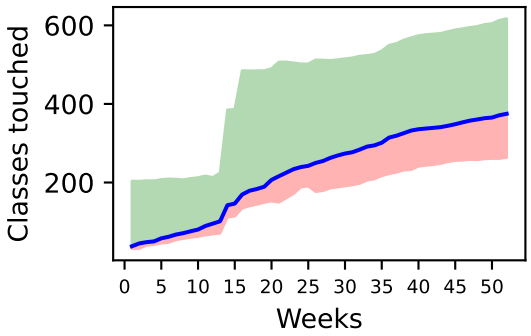
(c) Moderate founder developers (n=5)



(d) Moderate later joiner developers (n=85)



(e) Sustained founder developers (n=6)



(f) Sustained later joiner developers (n=21)

Figure 22: Repository: 14. A time series of the average (mean) total classes touched (y-axis) against the number of week (x-axis) for six (6) categories of developer.

2.3 Time series components touched - 14 For methods touched for each period

A time series of methods touched on average each month.

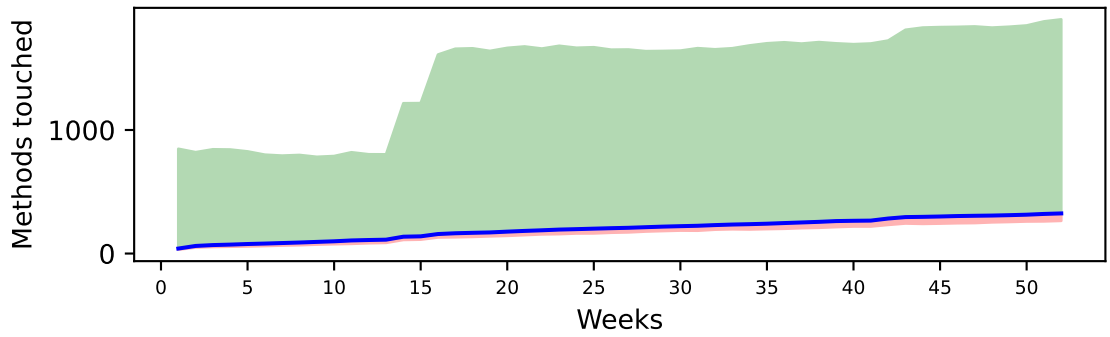
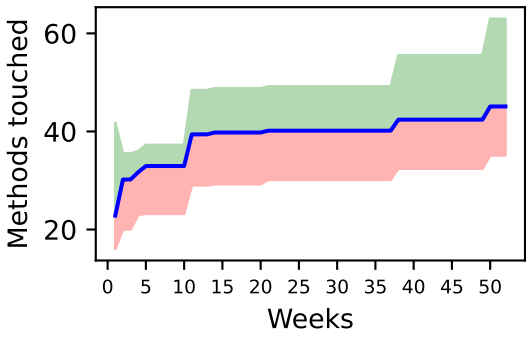
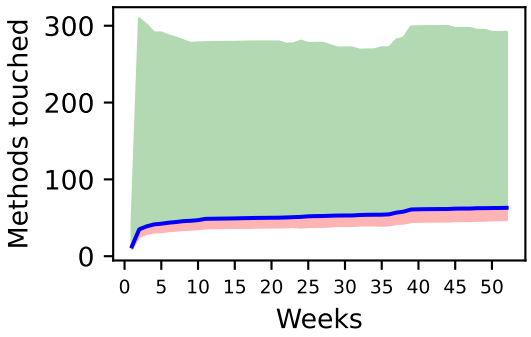


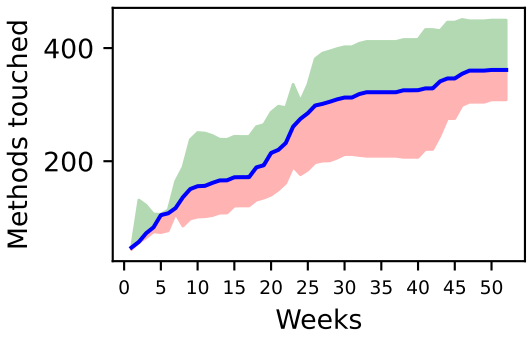
Figure 23: Repository: 14. A time series of the average (mean) total methods touched (y-axis) against the number of weeks (x-axis) for all developers (n=154).



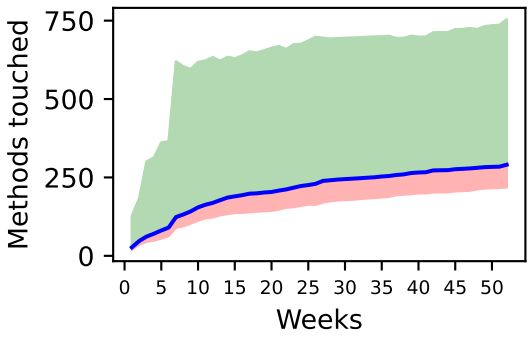
(a) Transient founder developers (n=8)



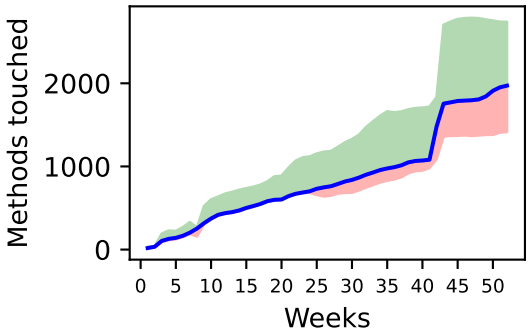
(b) Transient later joiner developers (n=119)



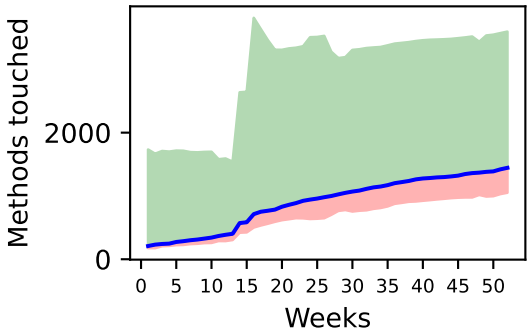
(c) Moderate founder developers (n=5)



(d) Moderate later joiner developers (n=85)



(e) Sustained founder developers (n=6)



(f) Sustained later joiner developers (n=21)

Figure 24: Repository: 14. A time series of the average (mean) total methods touched (y-axis) against the number of week (x-axis) for six (6) categories of developer.

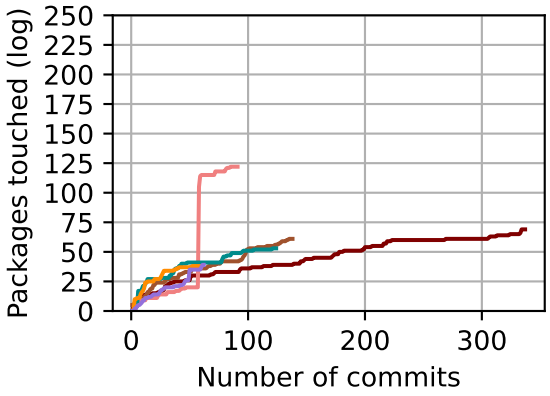
2.4 Time series sustained components touched by developer - 14 For developer commits and components touched.

Table 2: Table of first 6 sustained founder developers in the Scatter Developer graphs.

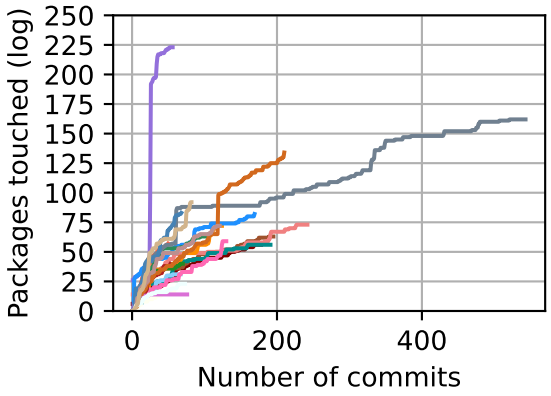
Developer ID	Colour
2930	Maroon
2932	Sienna
2934	LightCoral
2935	DarkCyan
2946	DarkOrange
2959	MediumPurple

Table 3: Table of first 21 sustained later joiner developers in the Scatter Developer graphs.

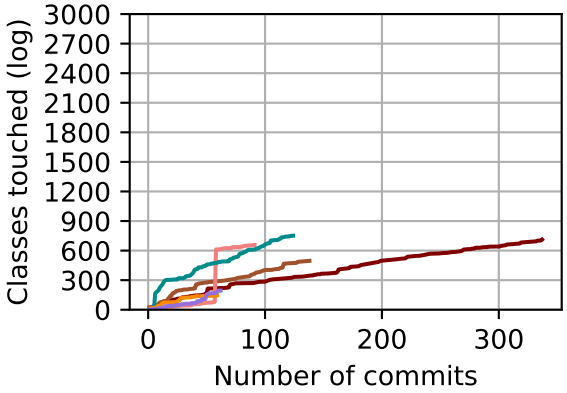
Developer ID	Colour
1774	Maroon
2986	Sienna
3000	LightCoral
3020	DarkCyan
3037	DarkOrange
3050	MediumPurple
3071	LightSkyBlue
3072	HotPink
3141	ForestGreen
3143	FireBrick
3163	DodgerBlue
3172	Tomato
3178	SeaGreen
3216	SlateGray
3217	Ivory
3272	Orchid
3339	Chocolate
3354	SteelBlue
3456	RosyBrown
3473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



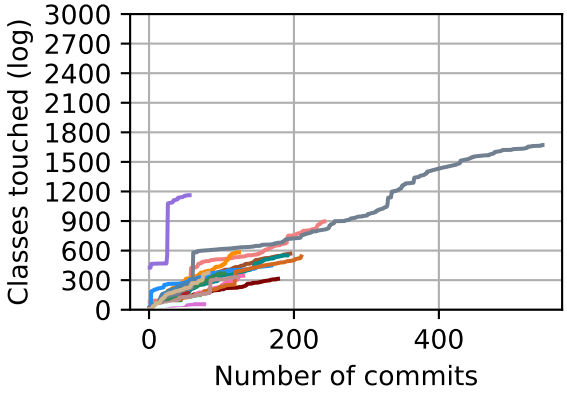
(a) Packages for sustained founder



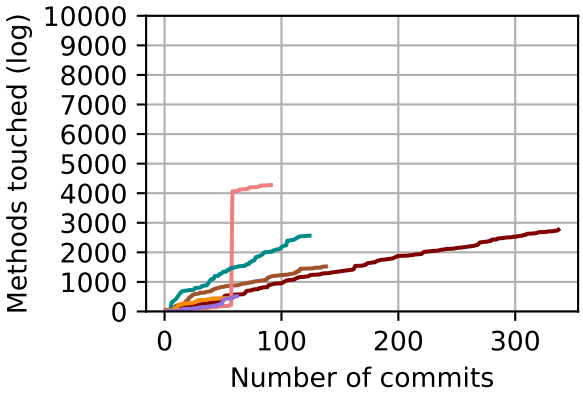
(b) Packages for sustained later joiner



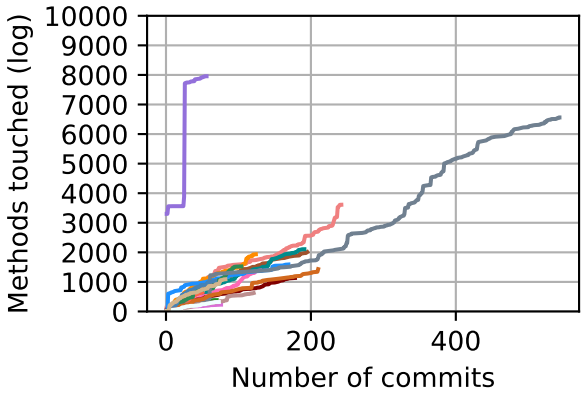
(c) Classes for sustained founder



(d) Classes for sustained later joiner



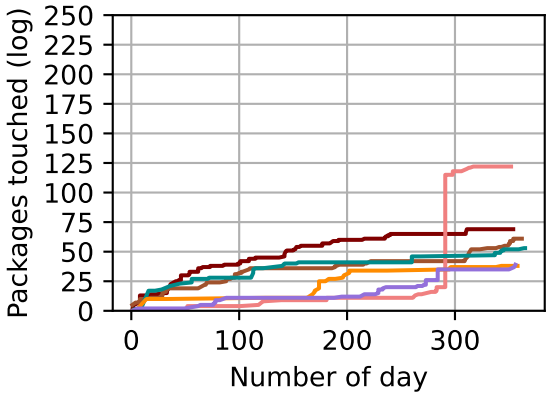
(e) Methods for sustained founder



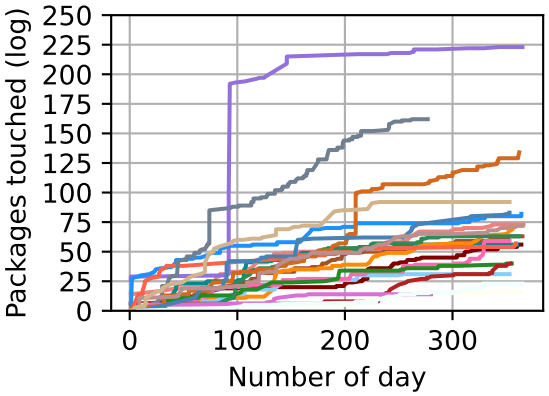
(f) Methods for sustained later joiner

Figure 25: Repo: 14. First six (6) sustained founder and first twenty-one (21) sustained later joiner developers. The number of components touched (y-axis) against the number of commits (x-axis). Sustained founder developers with colours are shown in Table 2. Sustained later joiner developers with colours are shown in Table 3.

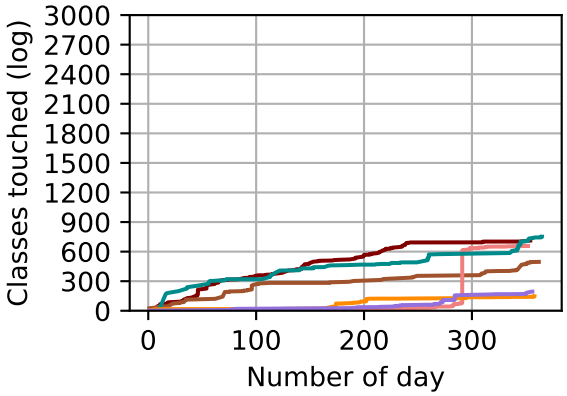
2.5 Time series sustained components touched by developer - 14 For developer day and components touched.



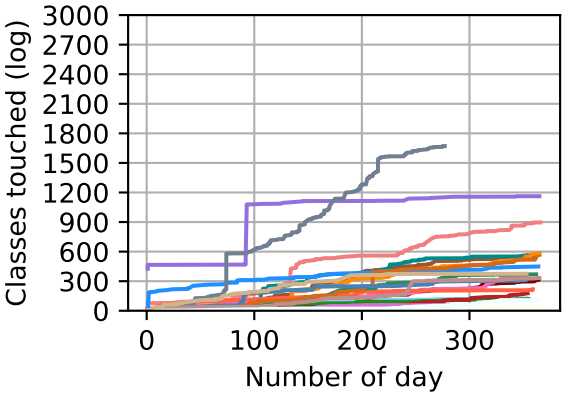
(a) Packages for sustained founder



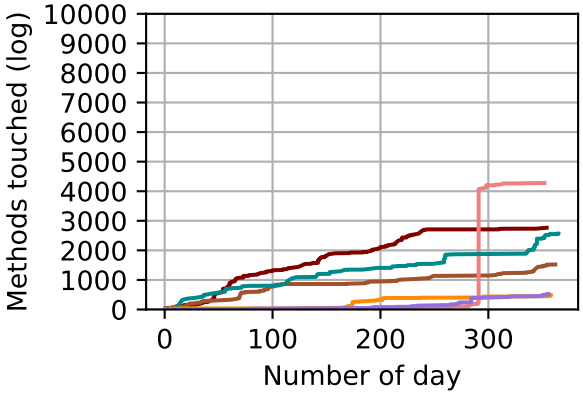
(b) Packages for sustained later joiner



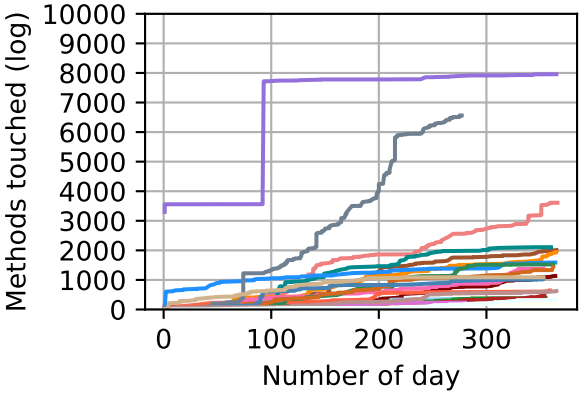
(c) Classes for sustained founder



(d) Classes for sustained later joiner



(e) Methods for sustained founder



(f) Methods for sustained later joiner

Figure 26: Repo: 14. First six (6) sustained founder and first twenty-one (21) sustained later joiner developers. The number of components touched (y-axis) against the number of day (x-axis). Sustained founder developers with colours are shown in Table 2. Sustained later joiner 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.

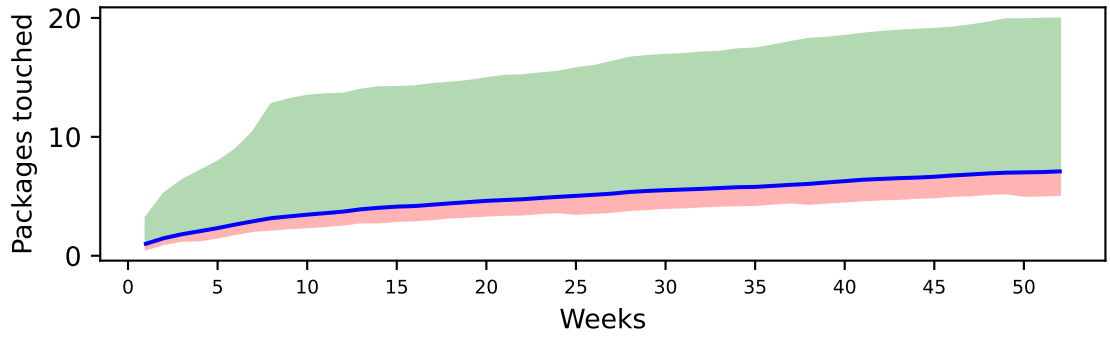
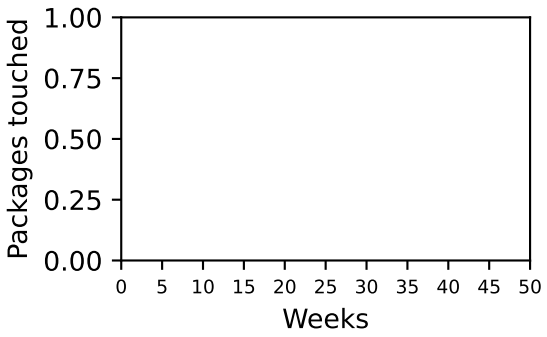
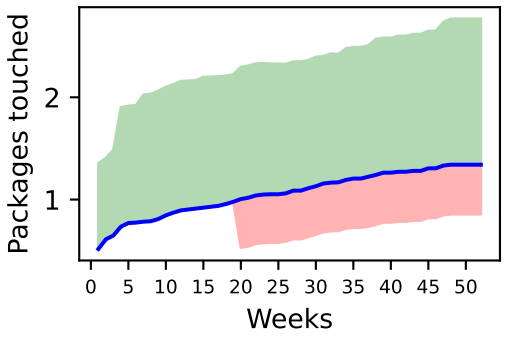


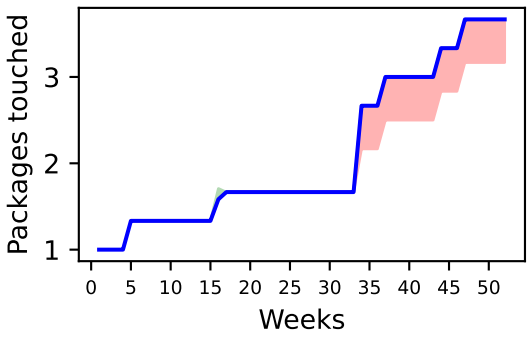
Figure 27: Repository: 21. A time series of the average (mean) total packages touched (y-axis) against the number of weeks (x-axis) for all developers (n=204).



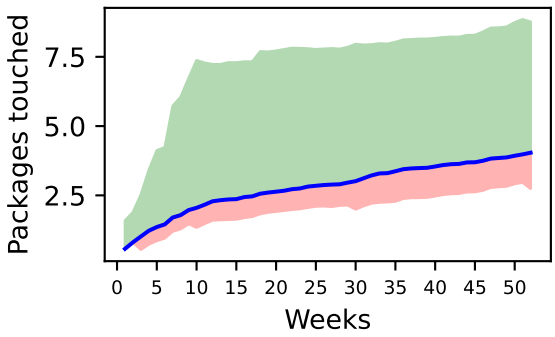
(a) Transient founder developers (n=0)



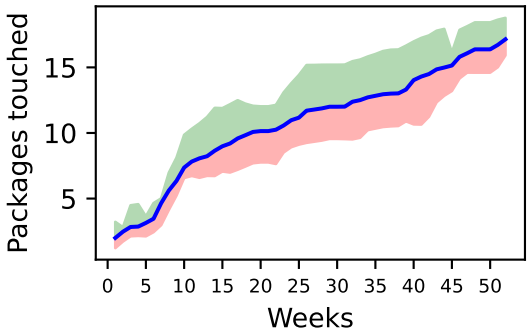
(b) Transient later joiner developers (n=113)



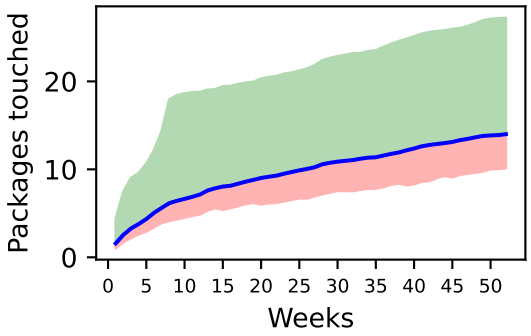
(c) Moderate founder developers (n=3)



(d) Moderate later joiner developers (n=93)



(e) Sustained founder developers (n=7)



(f) Sustained later joiner developers (n=84)

Figure 28: Repository: 21. A time series of the average (mean) total packages touched (y-axis) against the number of week (x-axis) for six (6) categories of developer.

3.2 Time series components touched - 21 For classes touched for each period

A time series of classes touched on average each month.

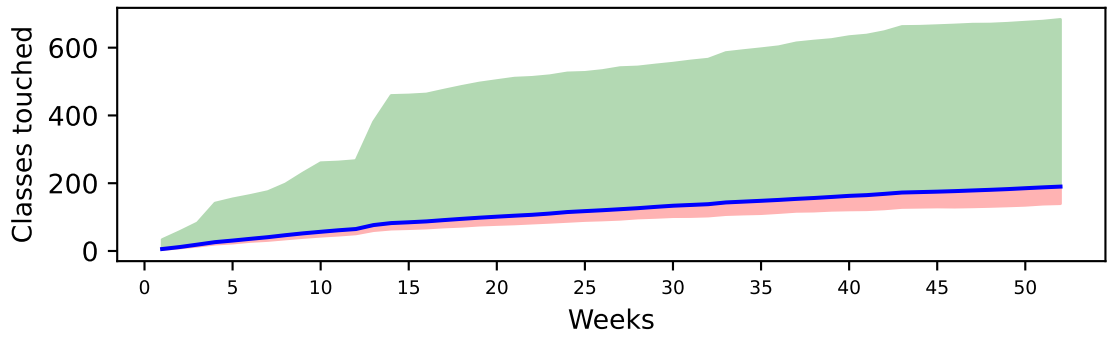
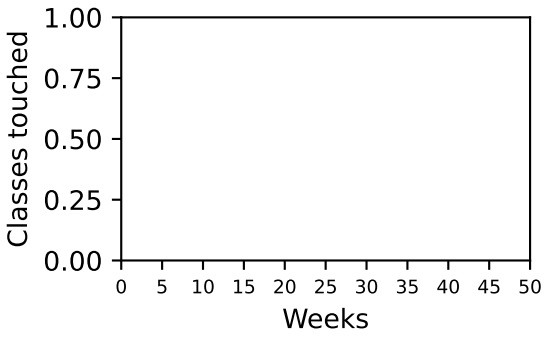
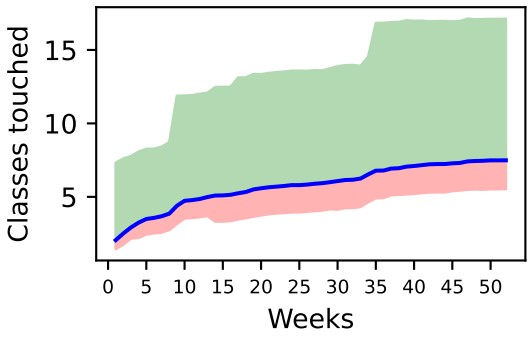


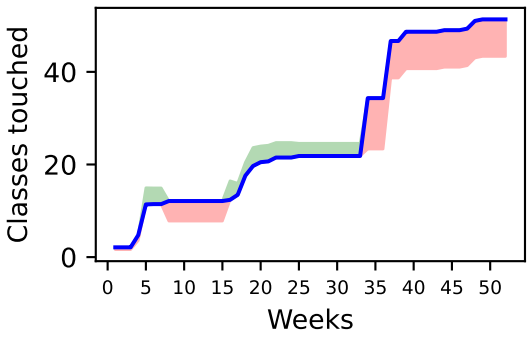
Figure 29: Repository: 21. A time series of the average (mean) total classes touched (y-axis) against the number of weeks (x-axis) for all developers (n=204).



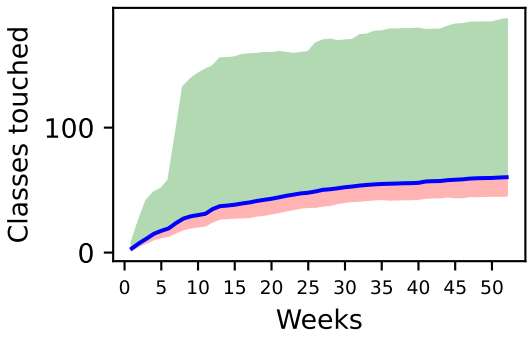
(a) Transient founder developers (n=0)



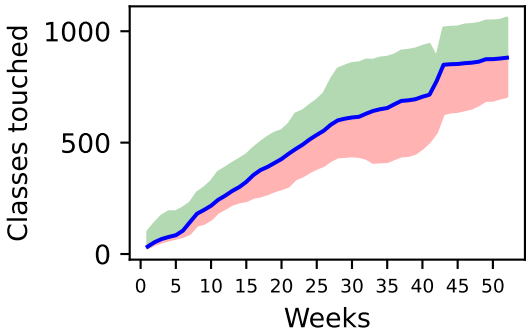
(b) Transient later joiner developers (n=113)



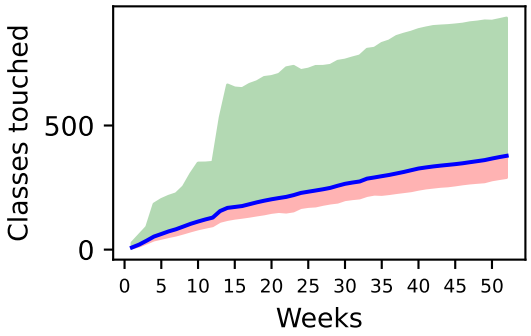
(c) Moderate founder developers (n=3)



(d) Moderate later joiner developers (n=93)



(e) Sustained founder developers (n=7)



(f) Sustained later joiner developers (n=84)

Figure 30: Repository: 21. A time series of the average (mean) total classes touched (y-axis) against the number of week (x-axis) for six (6) categories of developer.

3.3 Time series components touched - 21 For methods touched for each period

A time series of methods touched on average each month.

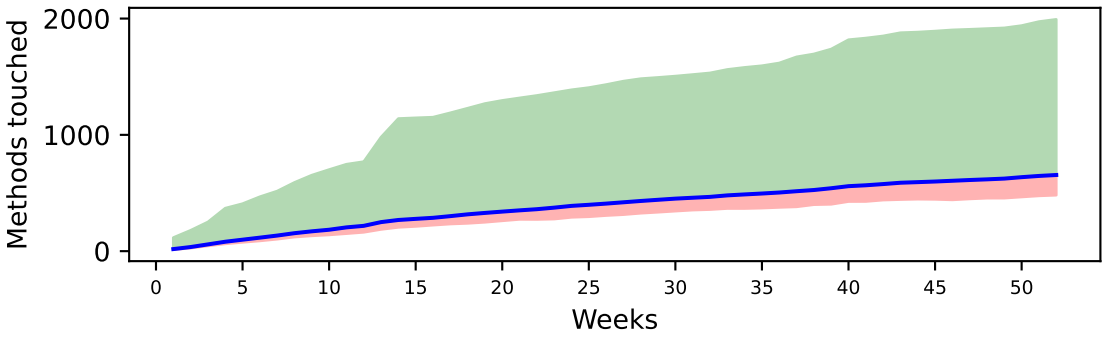
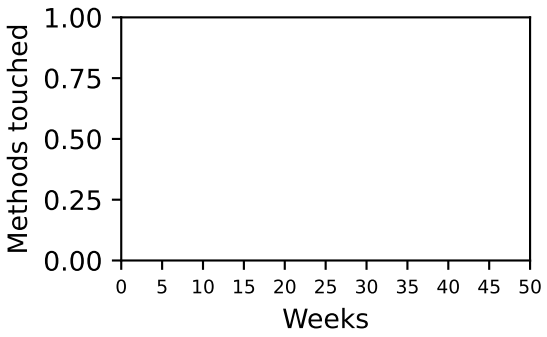
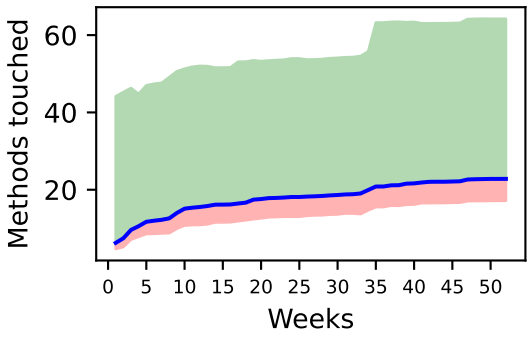


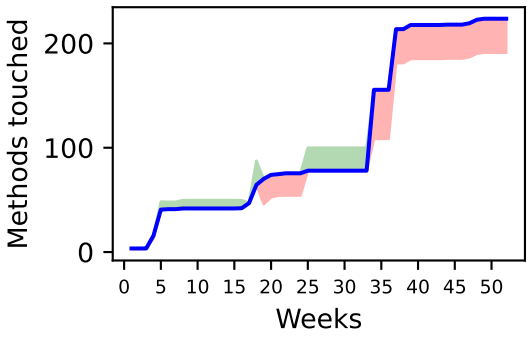
Figure 31: Repository: 21. A time series of the average (mean) total methods touched (y-axis) against the number of weeks (x-axis) for all developers (n=204).



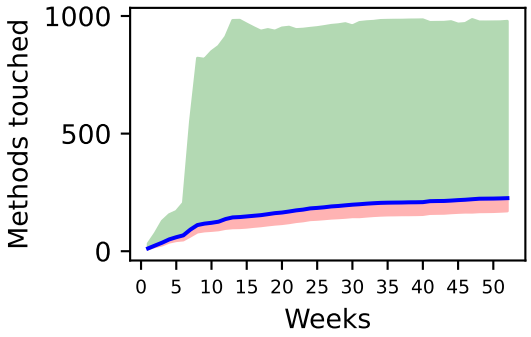
(a) Transient founder developers (n=0)



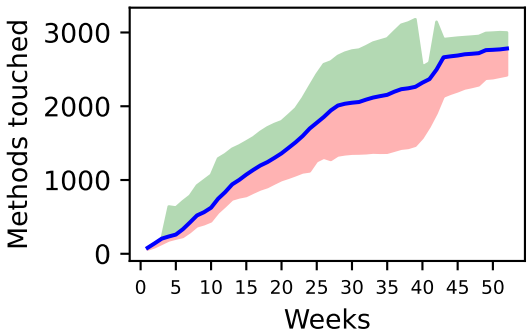
(b) Transient later joiner developers (n=113)



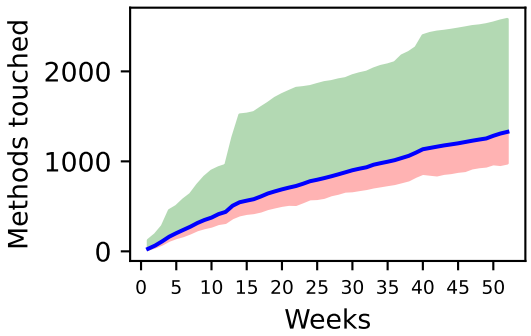
(c) Moderate founder developers (n=3)



(d) Moderate later joiner developers (n=93)



(e) Sustained founder developers (n=7)



(f) Sustained later joiner developers (n=84)

Figure 32: Repository: 21. A time series of the average (mean) total methods touched (y-axis) against the number of week (x-axis) for six (6) categories of developer.

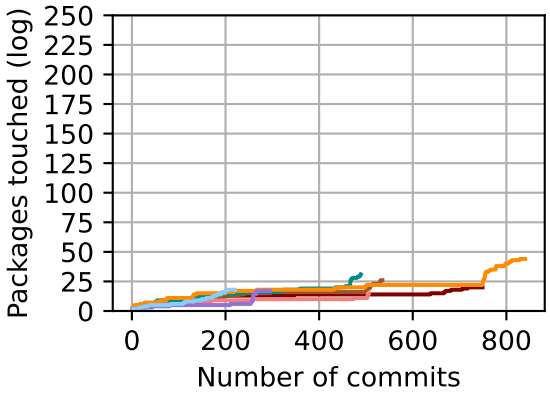
3.4 Time series sustained components touched by developer - 21 For developer commits and components touched.

Table 4: Table of first 7 sustained founder developers in the Scatter Developer graphs.

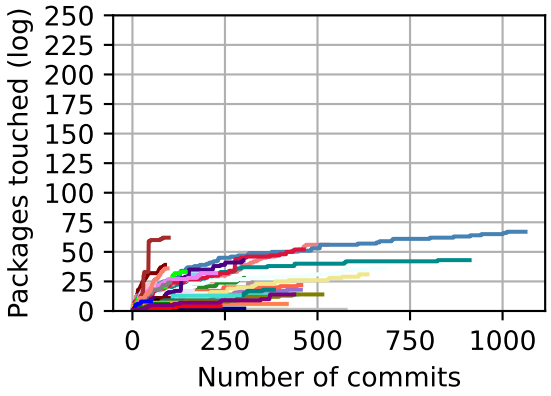
Developer ID	Colour
6245	Maroon
6246	Sienna
6247	LightCoral
6249	DarkCyan
6250	DarkOrange
6251	MediumPurple
6257	LightSkyBlue

Table 5: Table of first 50 sustained later joiner developers in the Scatter Developer graphs.

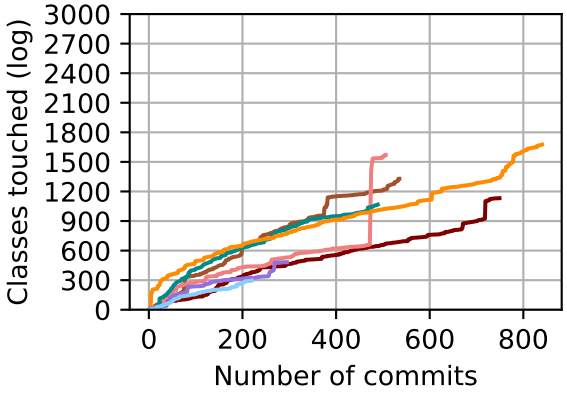
Developer ID	Colour
199	Maroon
6252	Sienna
6253	LightCoral
6259	DarkCyan
6260	DarkOrange
6262	MediumPurple
6263	LightSkyBlue
6264	HotPink
6267	ForestGreen
6271	FireBrick
6272	DodgerBlue
6276	Tomato
6278	SeaGreen
6280	SlateGray
6282	Ivory
6284	Orchid
6288	Chocolate
6295	SteelBlue
6297	RosyBrown
6298	Azure
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



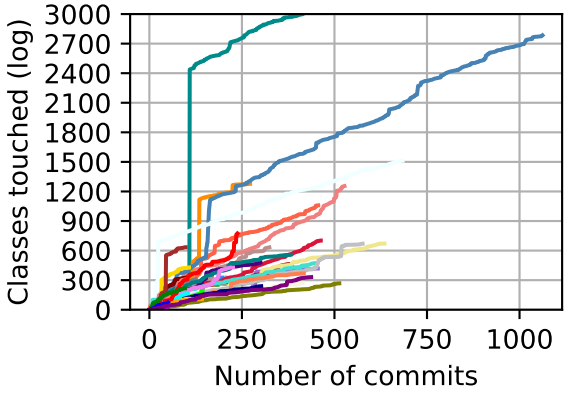
(a) Packages for sustained founder



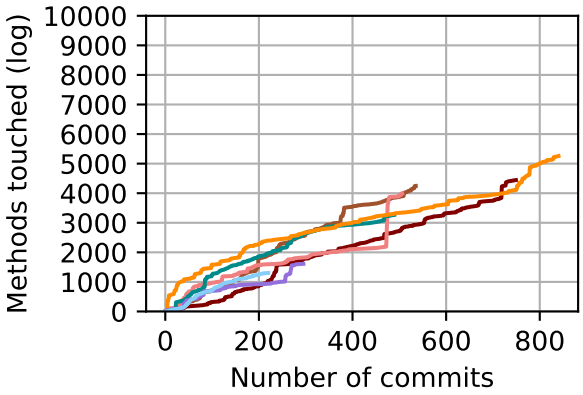
(b) Packages for sustained later joiner



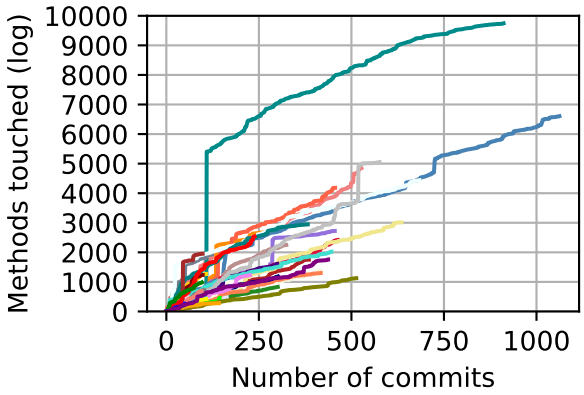
(c) Classes for sustained founder



(d) Classes for sustained later joiner



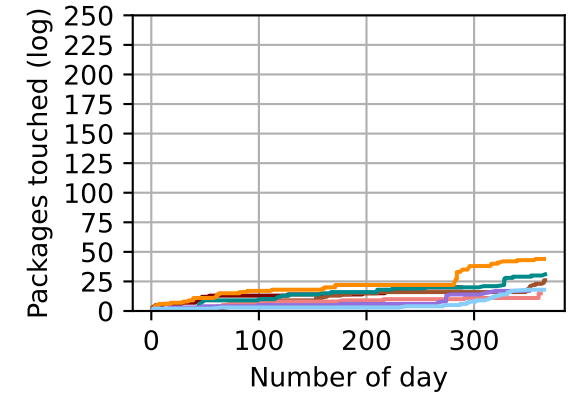
(e) Methods for sustained founder



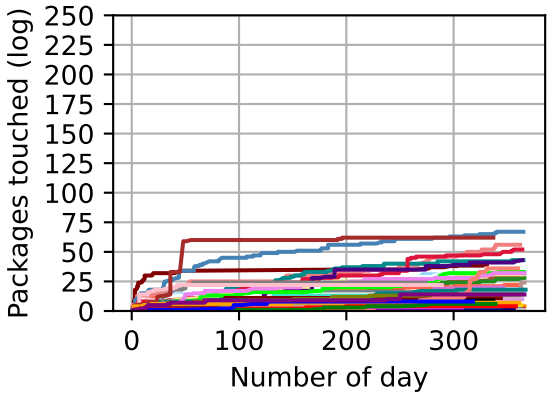
(f) Methods for sustained later joiner

Figure 33: Repo: 21. First seven (7) sustained founder and first fifty (50) sustained later joiner developers. The number of components touched (y-axis) against the number of commits (x-axis). Sustained founder developers with colours are shown in Table 4. Sustained later joiner developers with colours are shown in Table 5.

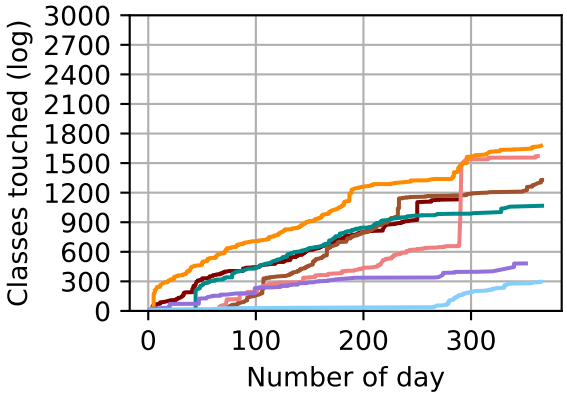
3.5 Time series sustained components touched by developer - 21 For developer day and components touched.



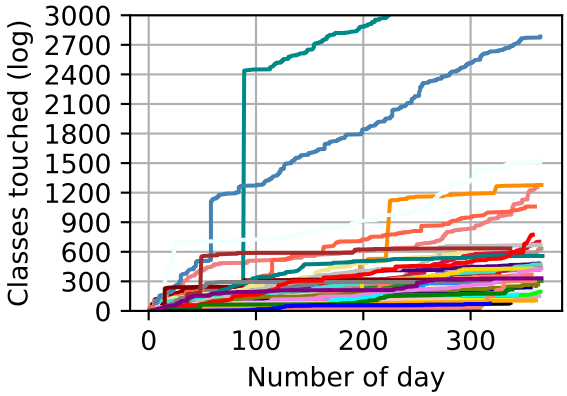
(a) Packages for sustained founder



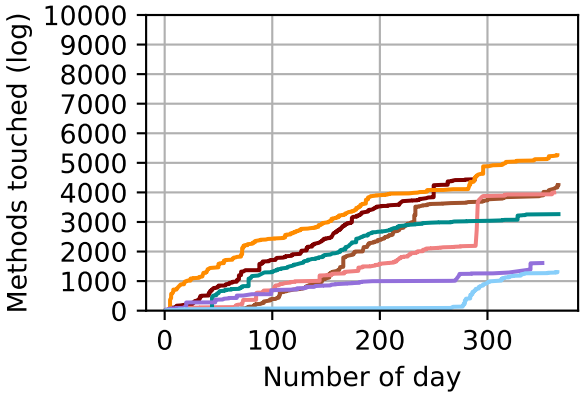
(b) Packages for sustained later joiner



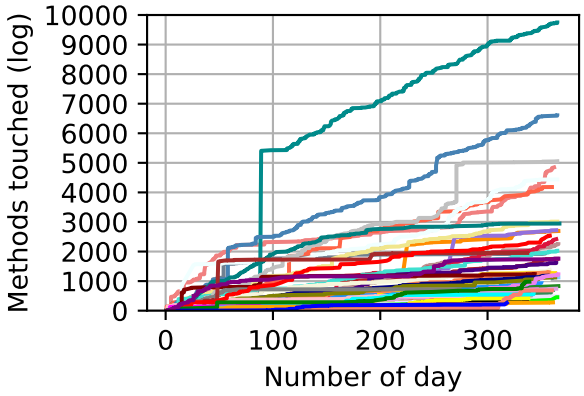
(c) Classes for sustained founder



(d) Classes for sustained later joiner



(e) Methods for sustained founder



(f) Methods for sustained later joiner

Figure 34: Repo: 21. First seven (7) sustained founder and first fifty (50) sustained later joiner developers. The number of components touched (y-axis) against the number of day (x-axis). Sustained founder developers with colours are shown in Table 4. Sustained later joiner developers with colours are shown in Table 5.