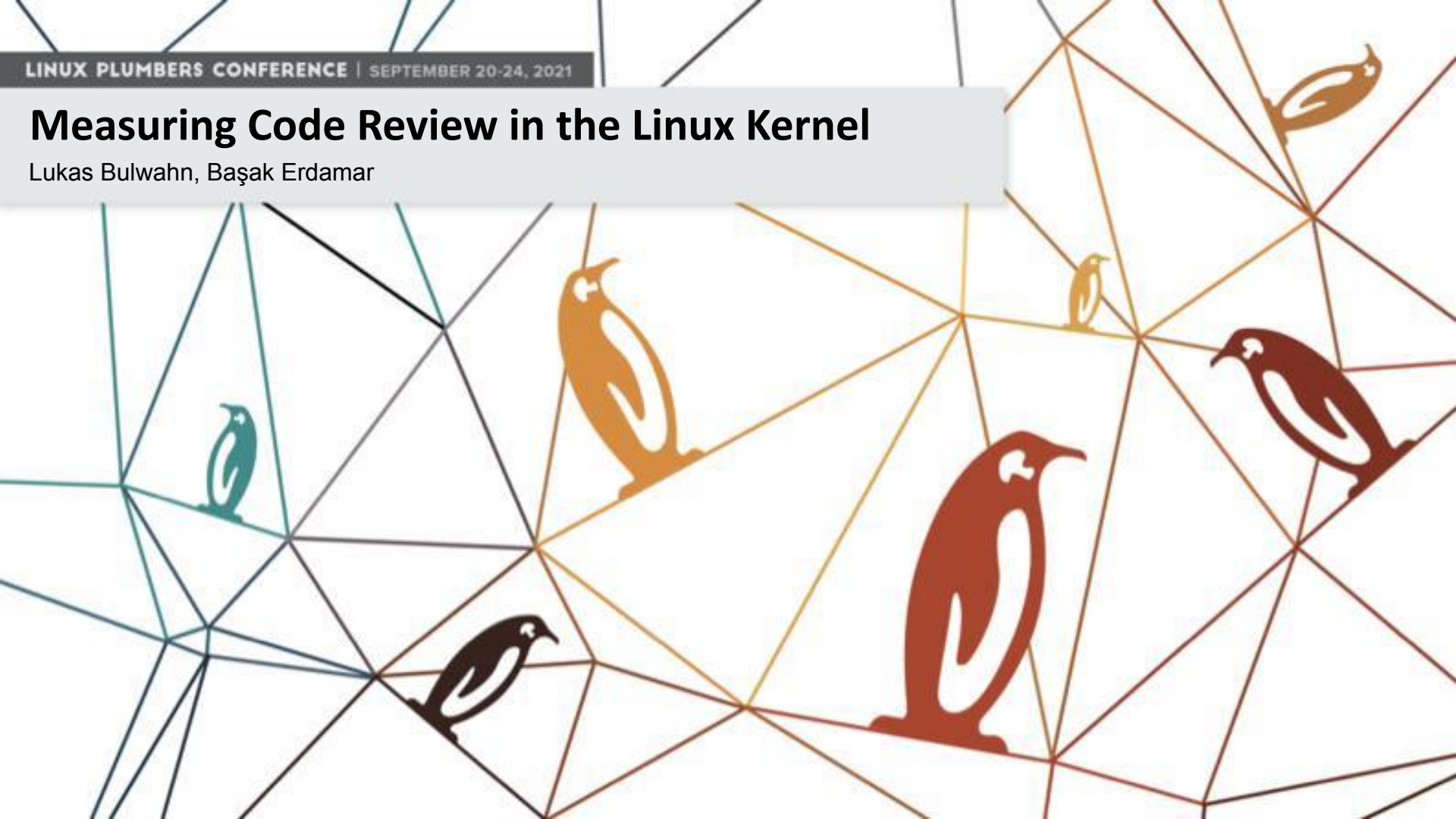
LINUX PLUMBERS CONFERENCE | SEPTEMBER 20-24, 2021

Measuring Code Review in the Linux Kernel

Lukas Bulwahn, Başak Erdamar



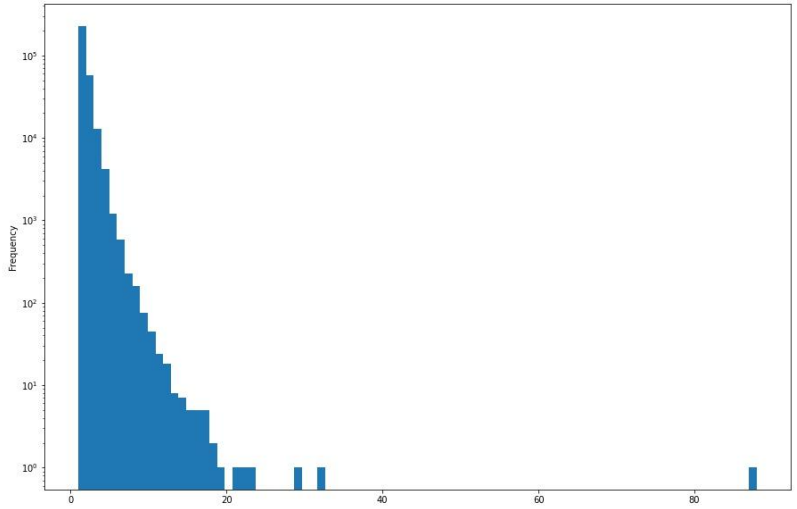
Introduction

Overview of Linux Kernel Development Process

- A patch is submitted to a relevant mailing list
- Patch is reviewed and integrated into associated git repository by the respective maintainer
- Commit is pulled from maintainer's integration tree and included in the main repository
- The potential problems about the patch is discovered through integration

Distribution of Number of Responses over Patches

- The focus for the thesis is the review that happens in the second step of the process.
- The aim is to measure the review and determine the factors contributing to it.
- The number of review emails is selected to be the central measure, self-responses are excluded.



Research Questions

- On patch authors:
  - Does the number of responses increase as the patch developer is more experienced?
  - Do maintainers get fewer or more responses than others, when they author a patch?
  - Do patch developers who have previously been active in some areas of the kernel get more responses than developers who have been active in other areas?
- On characteristics of the patches themselves:
  - Does the number of responses increase or decrease with the number of files a patch proposes to change?
  - Does the number of responses increase or decrease with the number of maintainer sections to which changed files belong to?
  - Does a patch get more responses if it is submitted to more mailing lists?
  - Do some mailing lists or maintainer sections lead to larger numbers of responses than others?

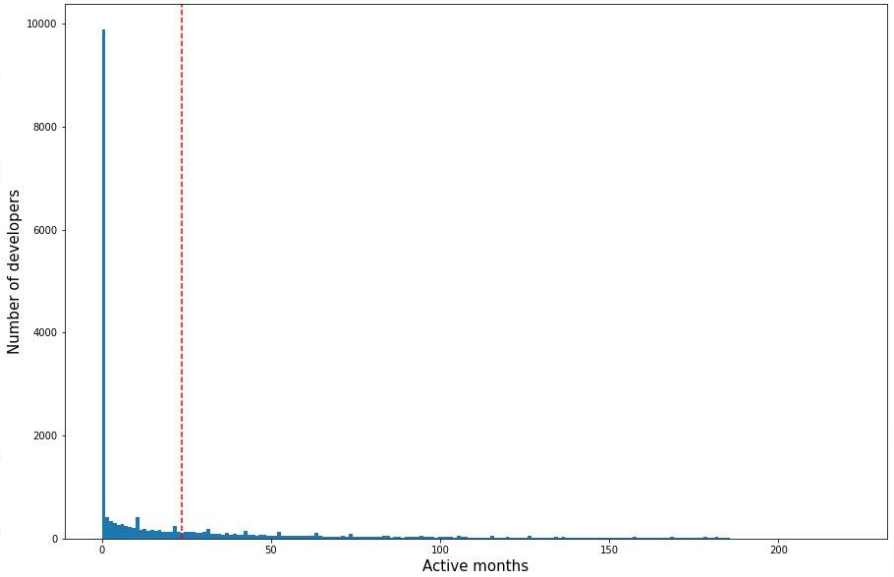
Topics

- Introduction
- Authoring Activity:
  - One Time Committers
- Maintainers
- Patches
- Bots

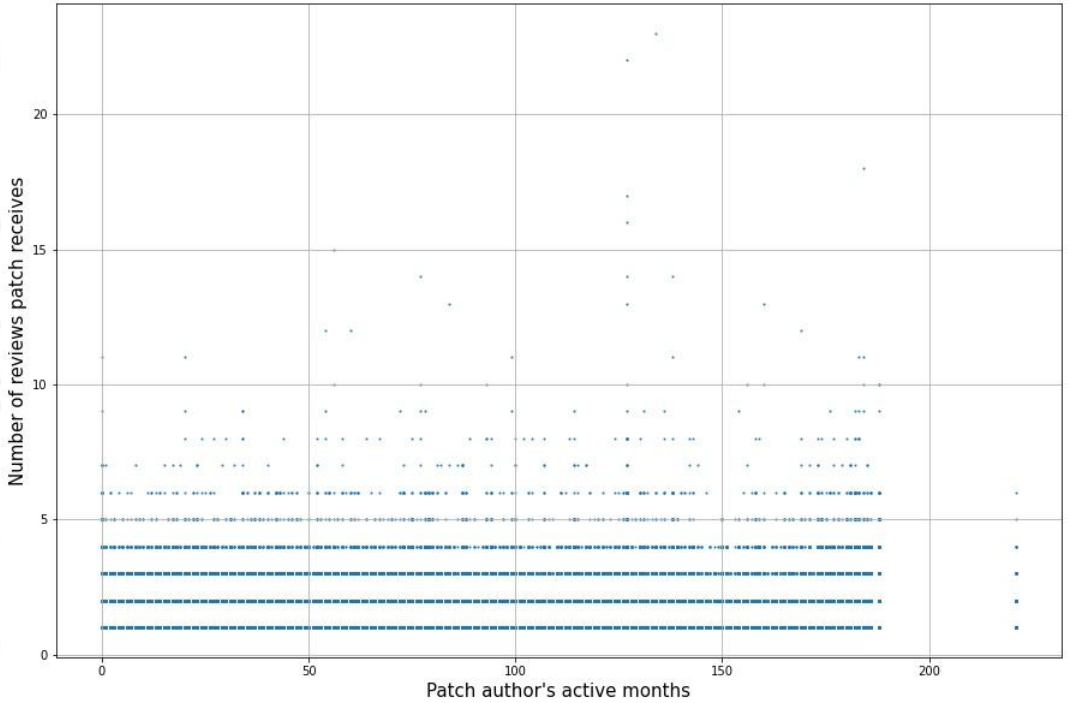
Authoring Activity

Authoring Activity: Active Months

- Many of the authors are relatively new in the kernel community.
- Having been in the community for a longer period of time shows no relation to the number of responses one's patch receives.

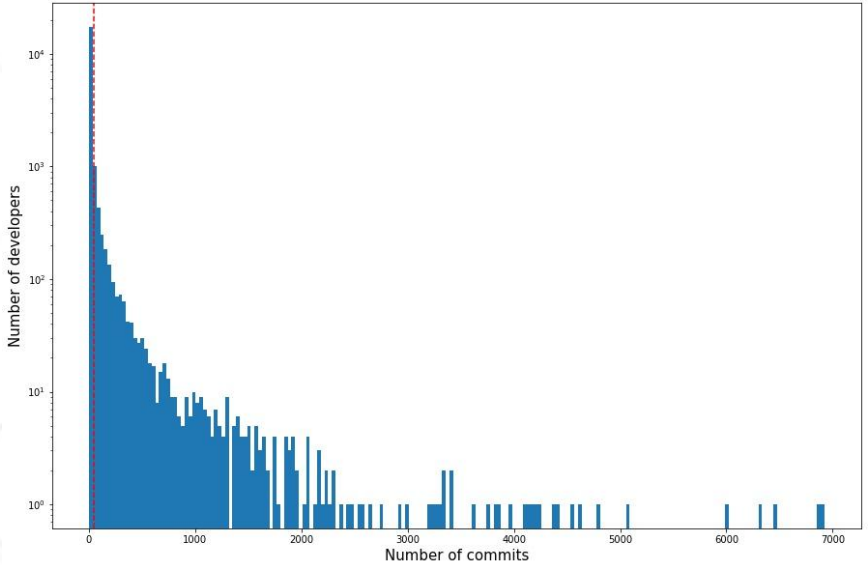


Authoring Activity: Active Months

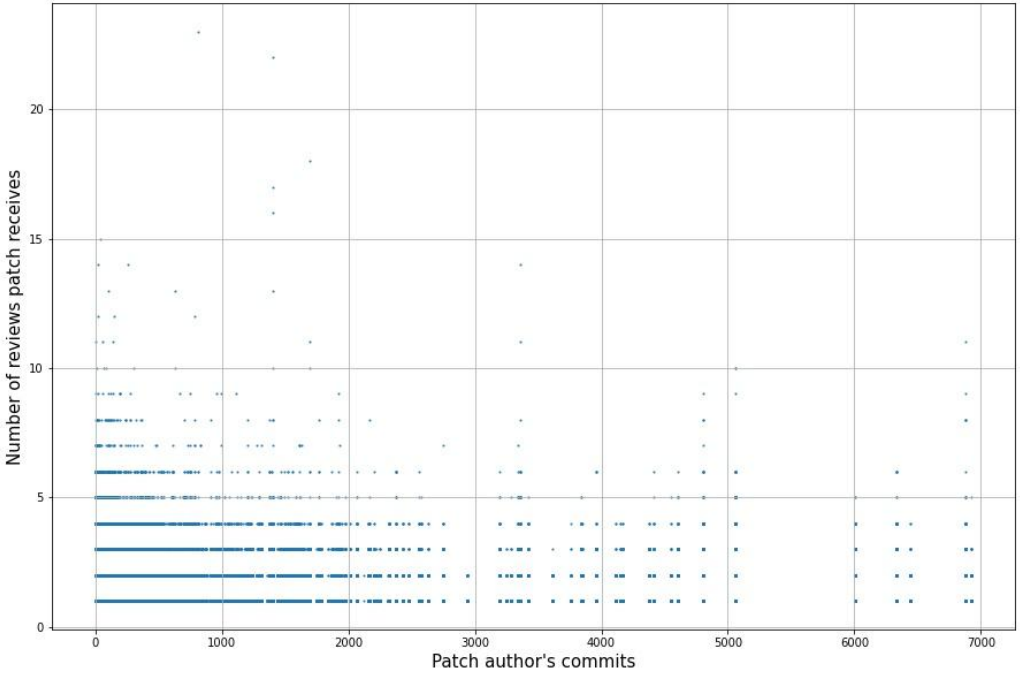


Authoring Activity: Commits

- An alternative measure for developer experience is the number of commits authored.
- Again, no clear positive relation to number of responses is seen.
- The histogram shows that one time committers to the kernel make up a large portion of the authors.

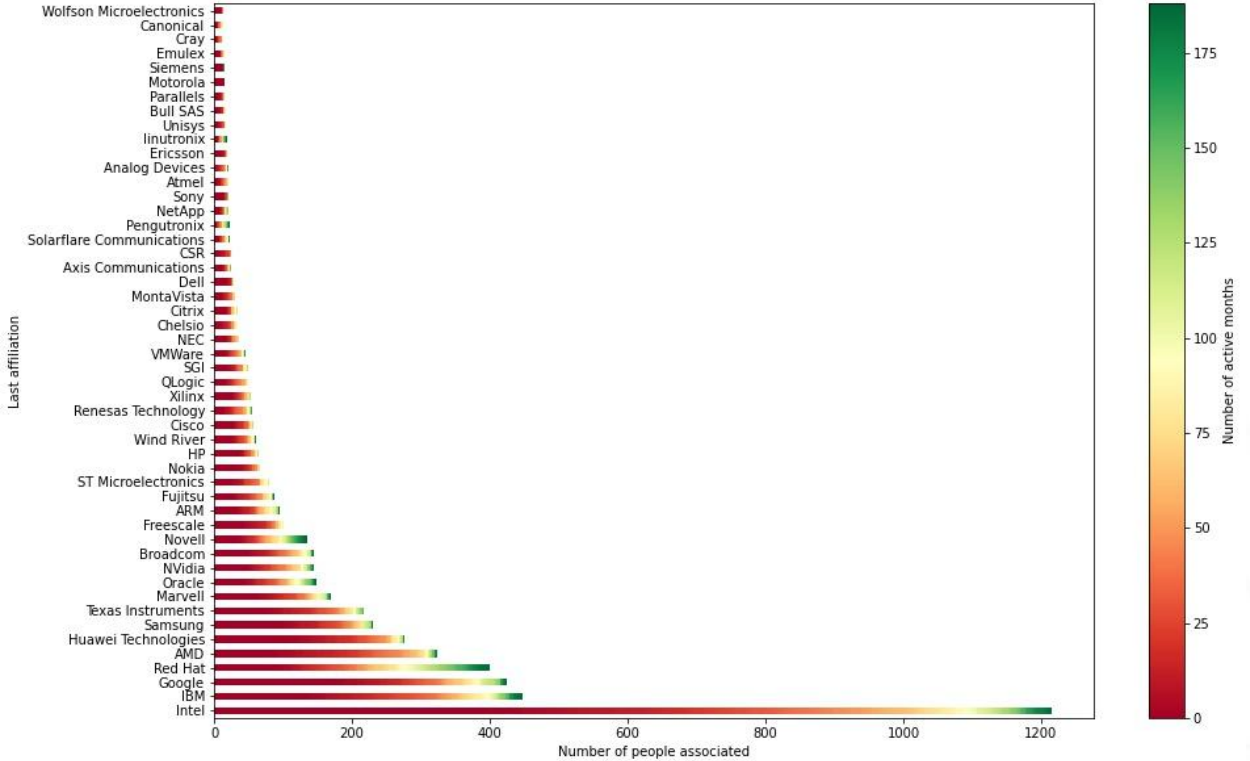


Authoring Activity: Commits



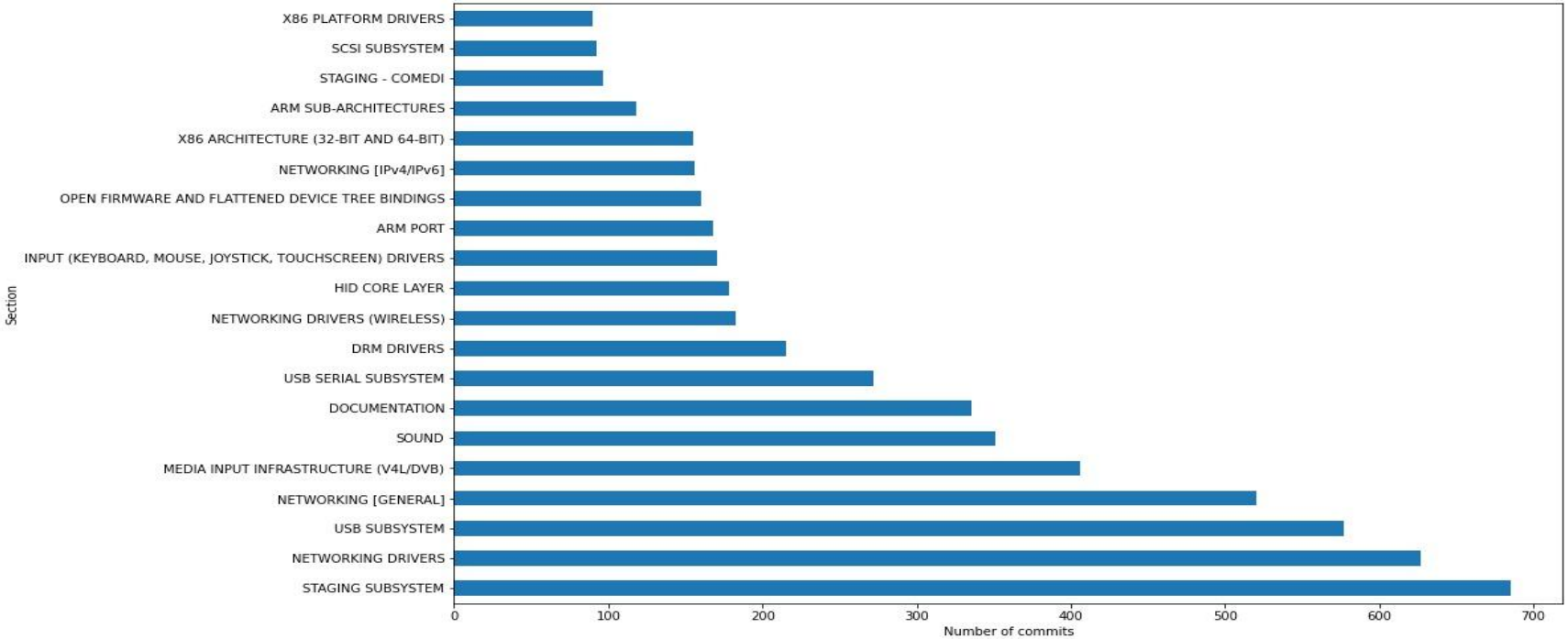
Authoring Activity: Top Companies

- Intel has 34.38% of its developers in the kernel having less than 2 active months.
- 76% Red Hat developers in the kernel has larger than 1 active months, 59% has larger that 24 months.

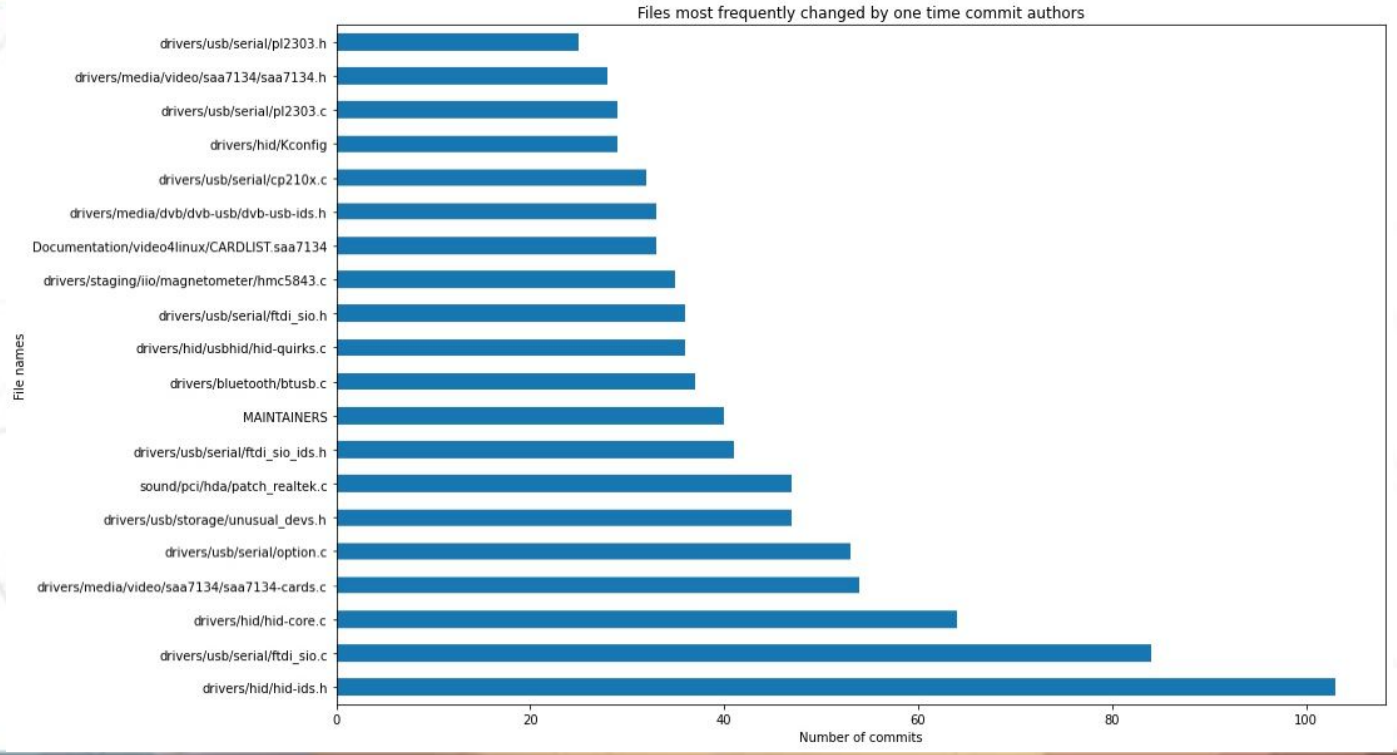


One Time Committers

One Time Committers' Popular Sections



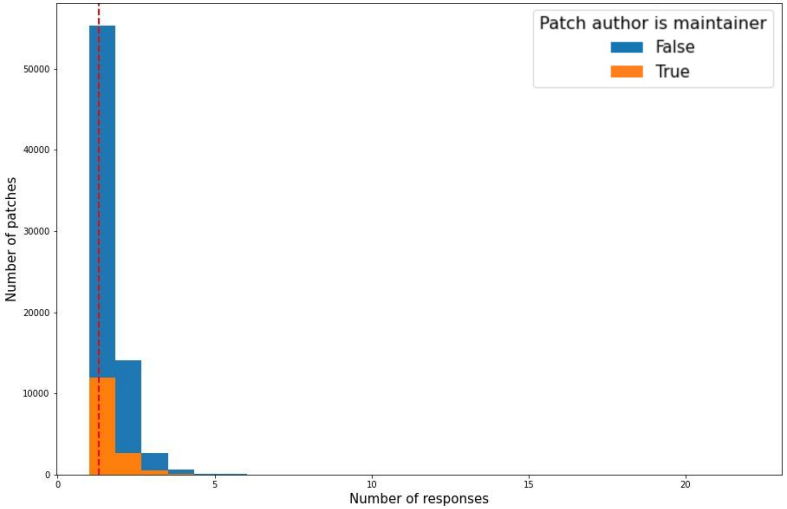
One Time Committers: Most Popular Files

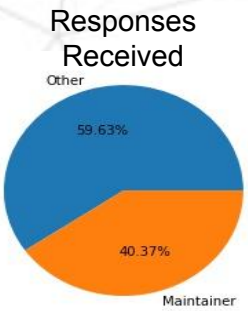


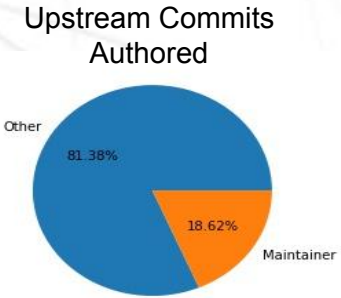
Maintainers

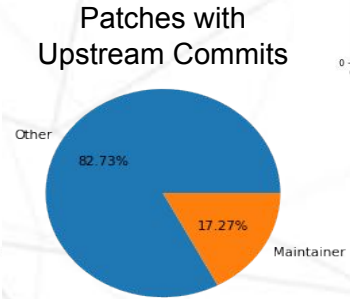
Maintainers: Authoring and Responses

- Maintainers make up 11.22% of the authors
- Maintainers received a larger portion of the responses than the portion of their authoring activity



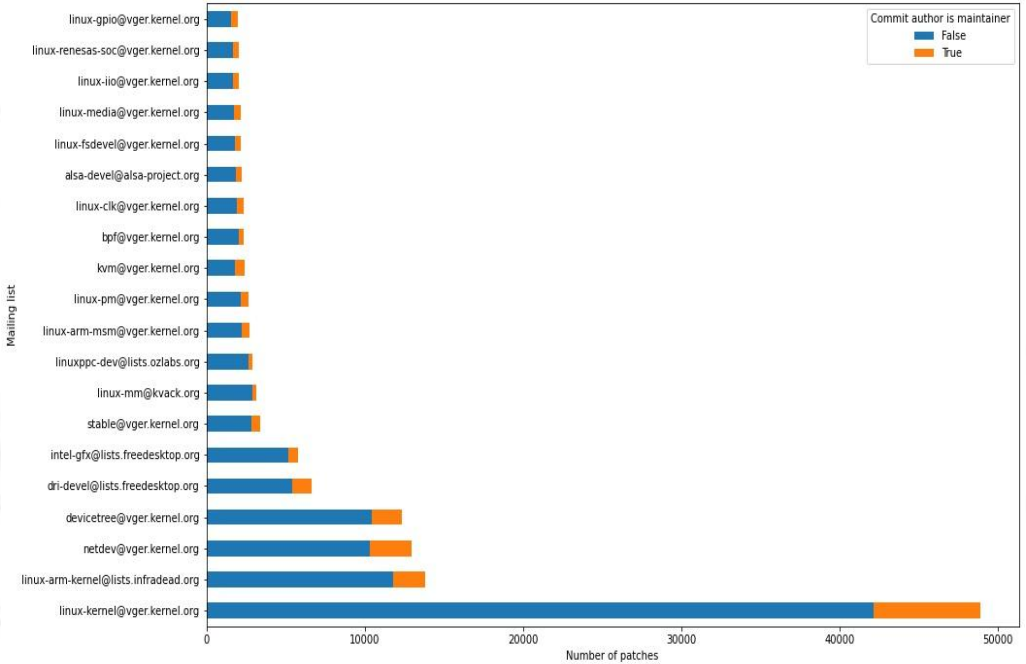






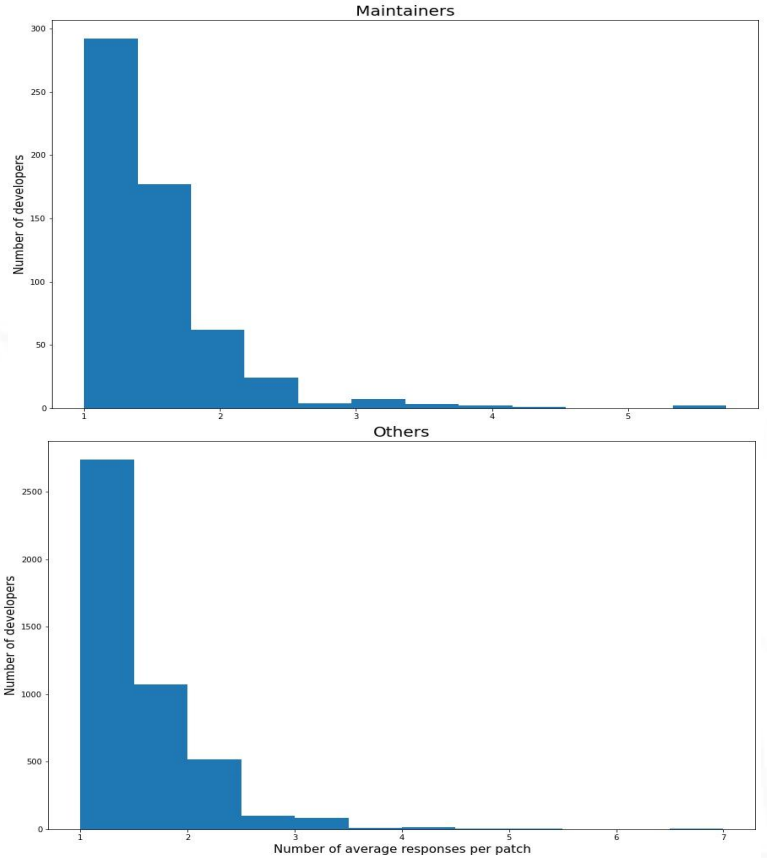
Maintainers: Top-20 Most Active Mailing Lists

- 13.81% percent of patches in linux-kernel@vger.kernel.org authored by maintainers
- The highest percentage of maintainer patches in the list is 25.98% seen on kvm@vger.kernel.org
- The lowest percentage of maintainer patches is 0% seen on 30 mailing lists.
- Among all of these top 20 most active lists, average percentage of maintainer patches is 16.69%



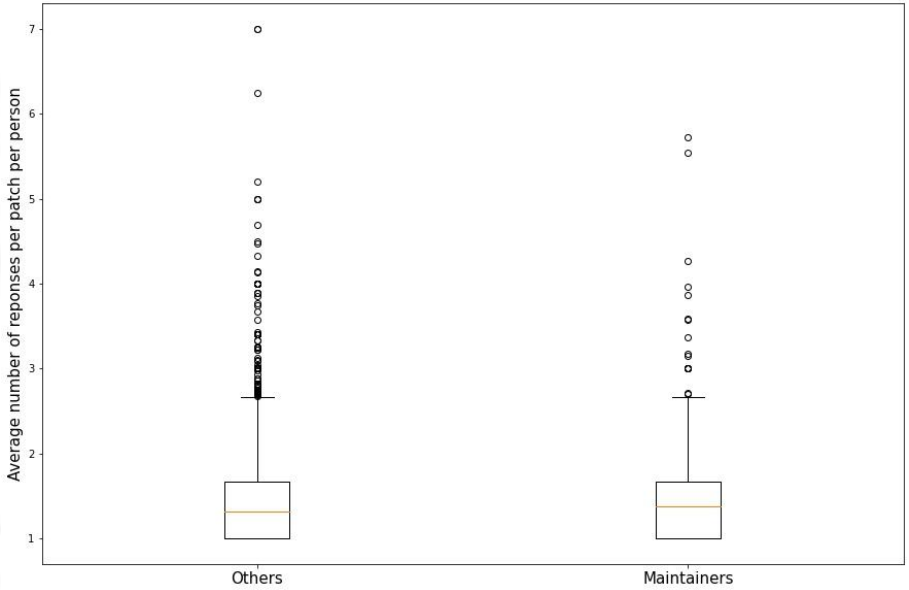
Average Number of Responses per Patch

- The histograms show distributions of average number of responses received per patch by maintainers and by others.
- The distributions for neither maintainers nor others are normal.
- Non-parametric rank sum test is be conducted to check the difference between maintainers and others



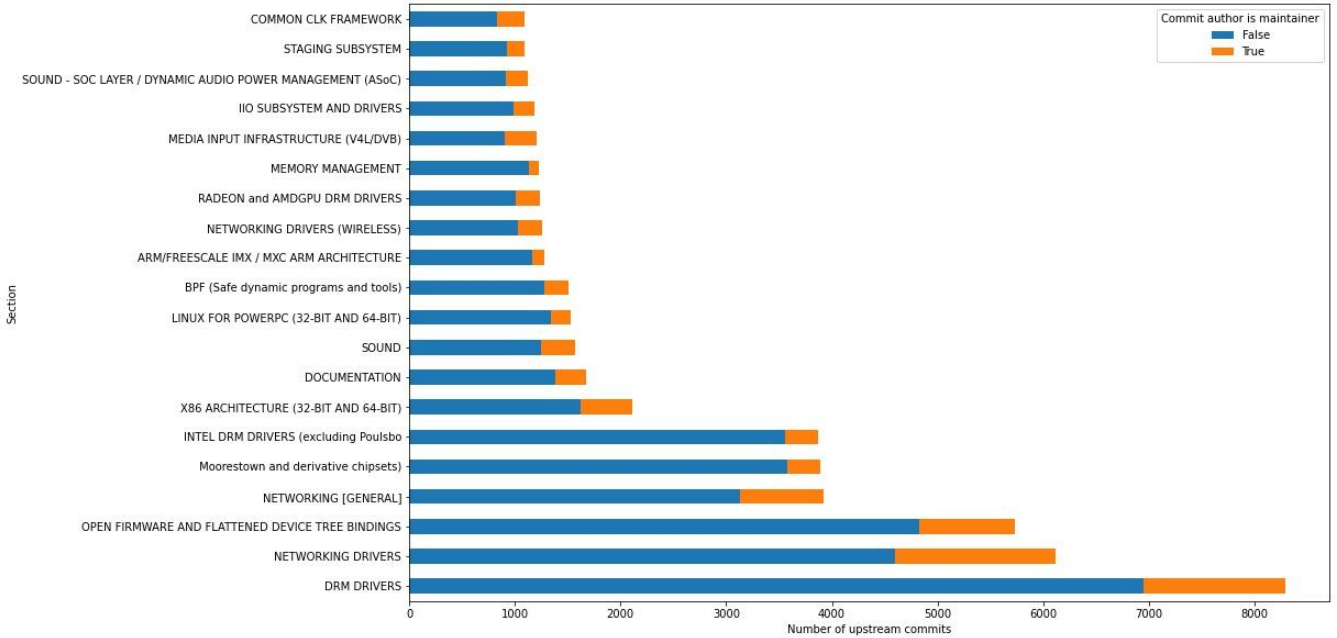
Average Number of Responses per Patch

- The Mann-Whitney test rejected the null hypothesis with a 95% confidence.
- There is enough evidence to reject that maintainers get the same average number of responses per patch



$$H_0 : \mu_{maintainers} = \mu_{others} \quad H_A : \mu_{maintainers} > \mu_{others}$$

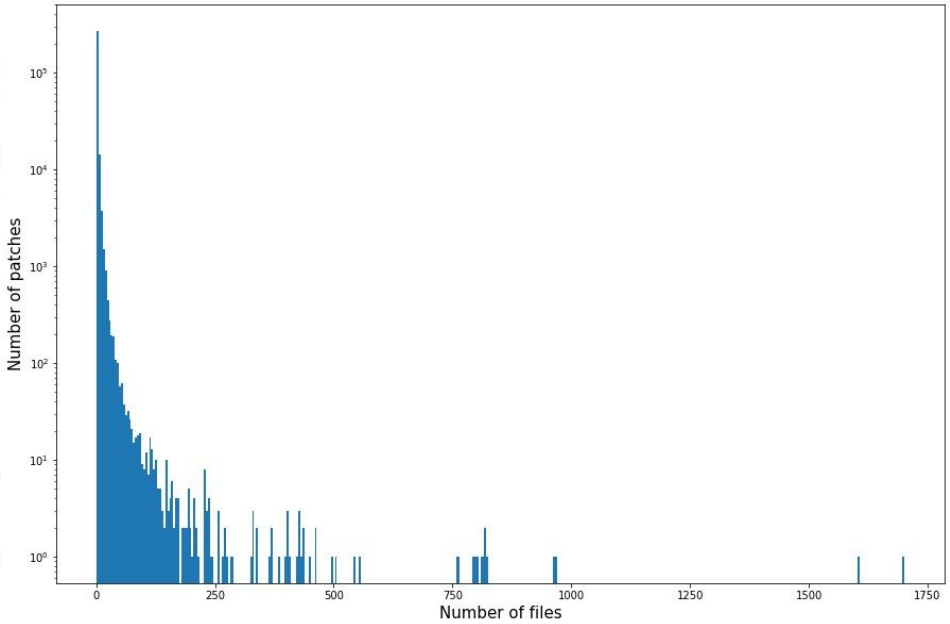
Maintainer Activity Across Sections



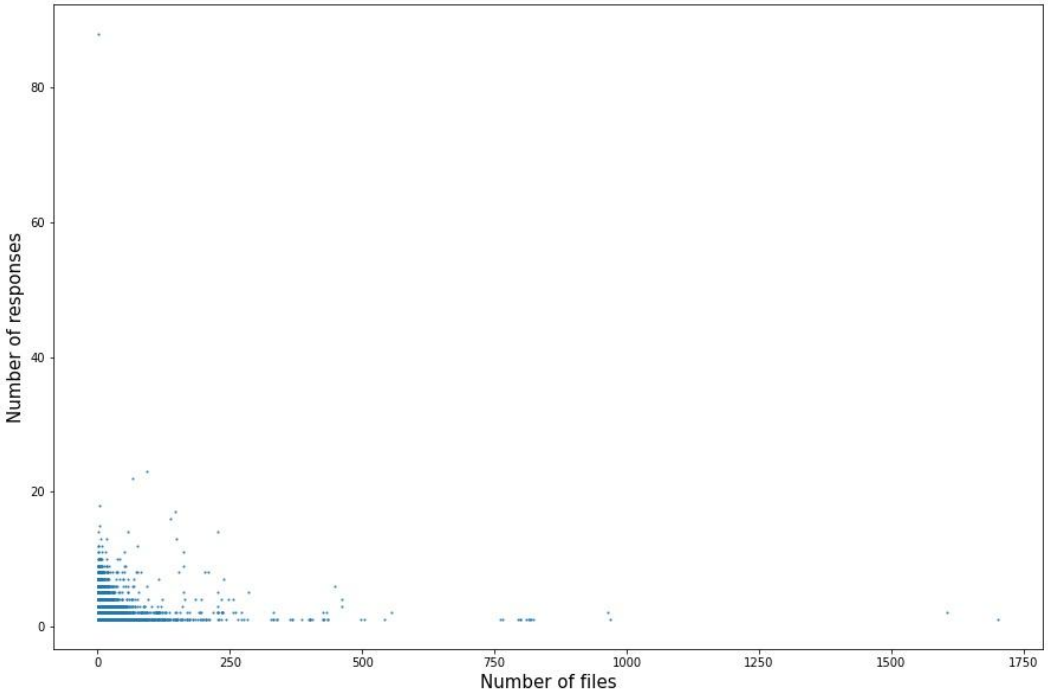
Patches

Patches: Number of Files

- Many patches change fewer files, only one in most cases
- While there are outliers, the number of files does not have a linear relation with the number of responses

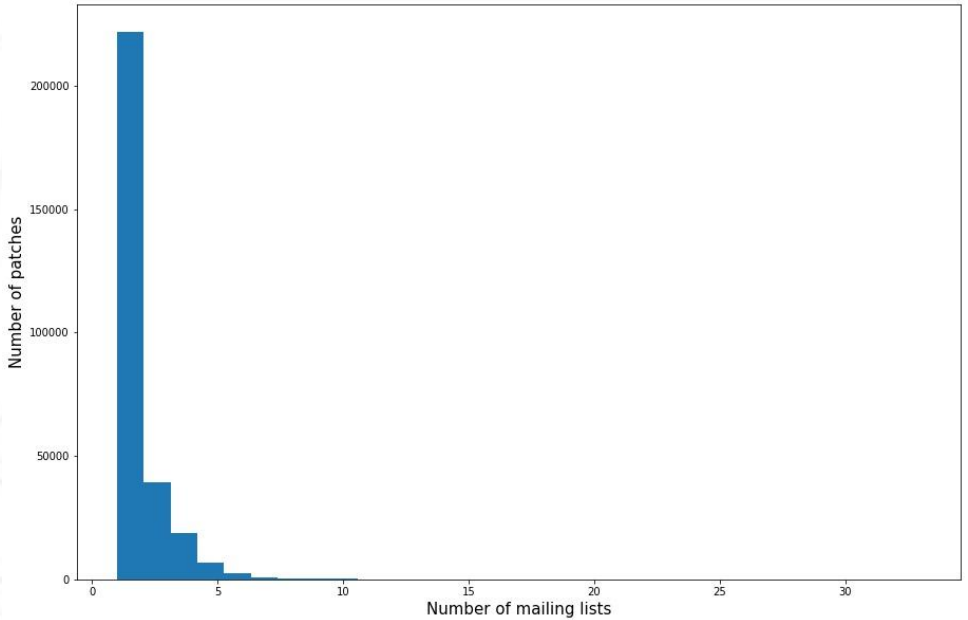


Patches: Number of Files

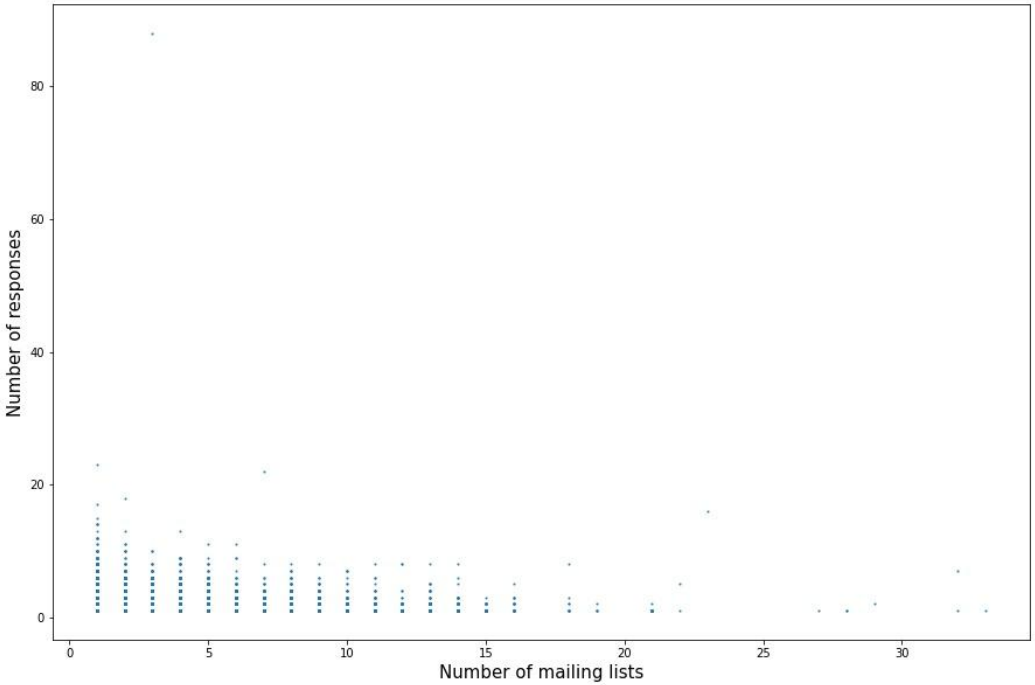


Patches: Number of Mailing Lists

- Next, whether sending a patch to more mailing lists result in more responses is inspected.
- No relation between overall number of mailing lists and the number of responses is seen
- What if we look at the effect of individual mailing lists instead of the total number of mailing lists?

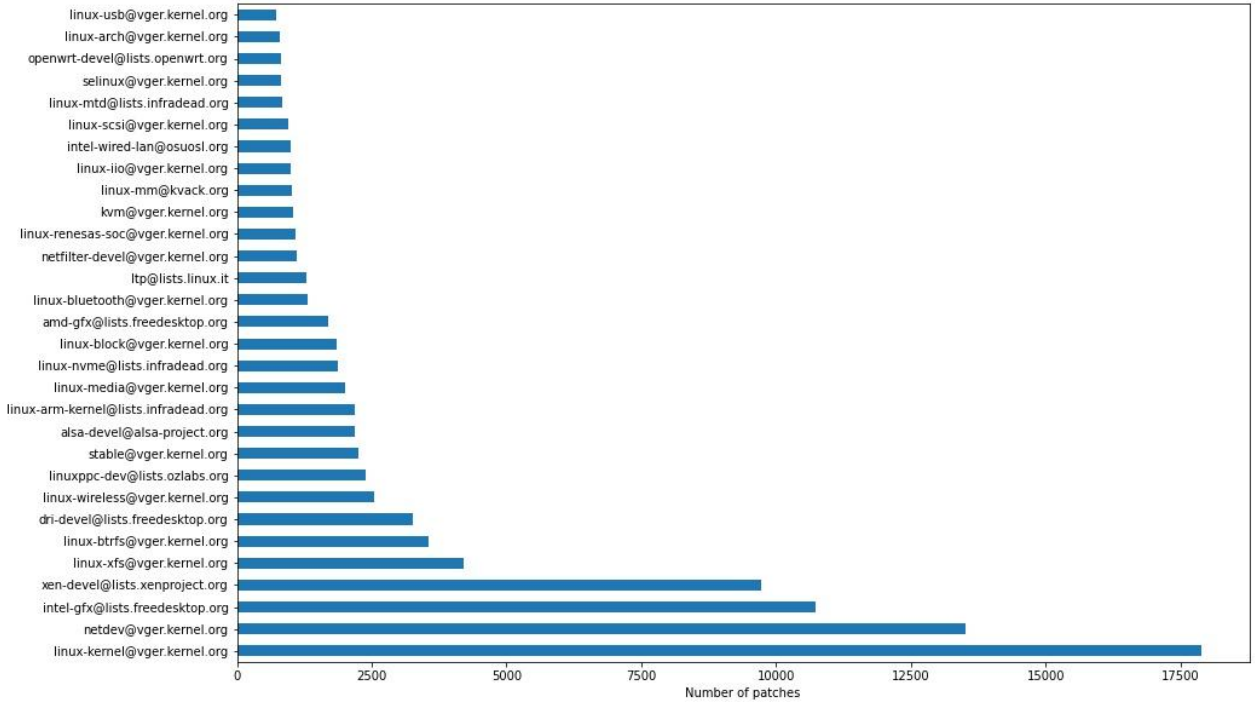


Patches: Number of Mailing Lists



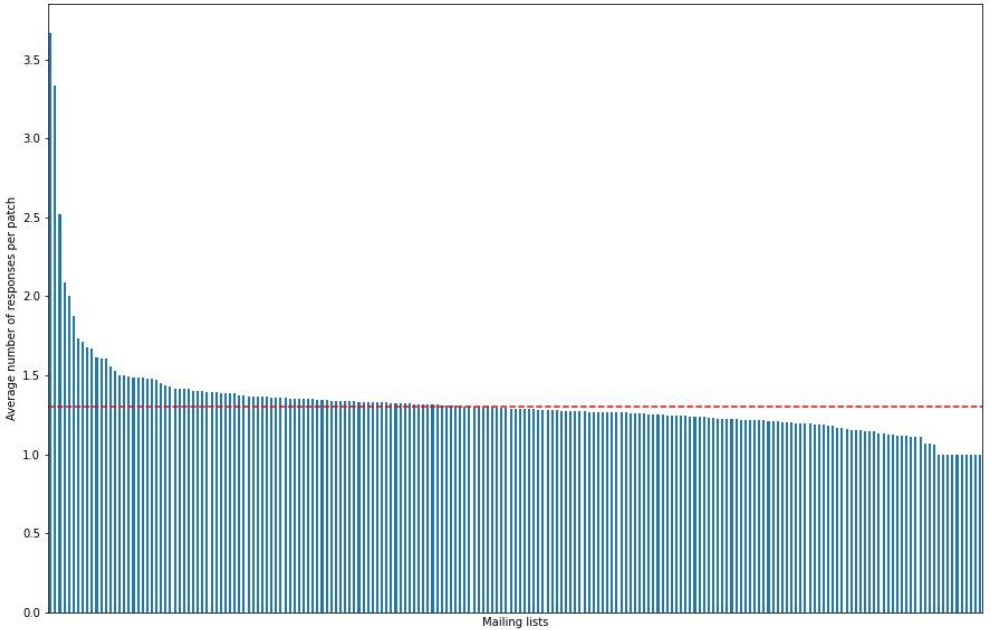
Isolated Mailing Lists

- 96.11% of all the emails that were sent to the leading list qemu-devel@nongnu.org were only sent to qemu-devel@nongnu.org.
- 45.48% of all the patches were submitted to linux-kernel@vger.kernel.org 15.16% of them were submitted only here.



Patches: Individual Mailing Lists

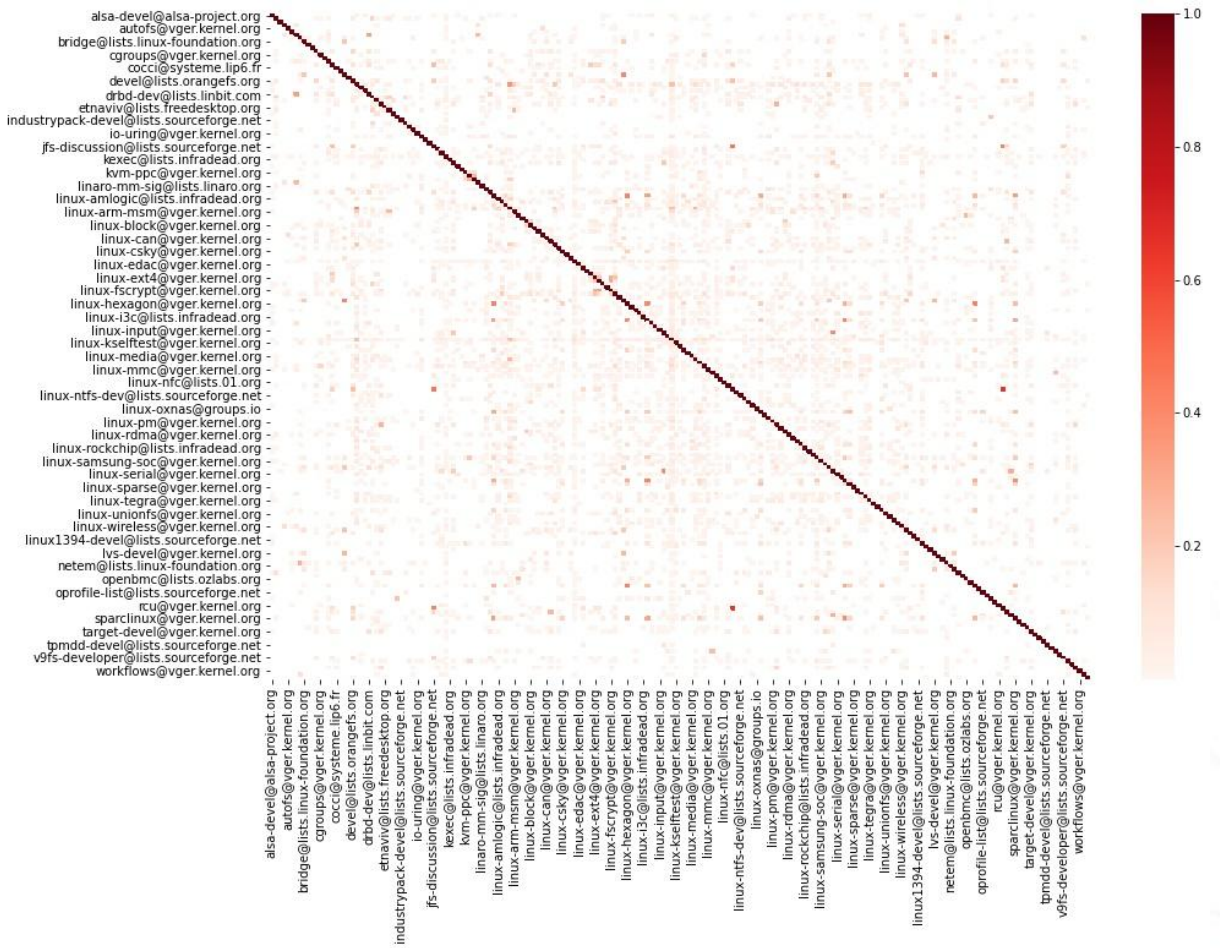
- The graph shows the average number of responses per patch for each of the mailing lists.
- The overall average number of responses per patch is 1.3.
- Patch itself is counted as one response
- Highest average is 3.67 on workflows@vger.kernel.org



LINUX PLUMBERS CONFERENCE September 20-24, 2021

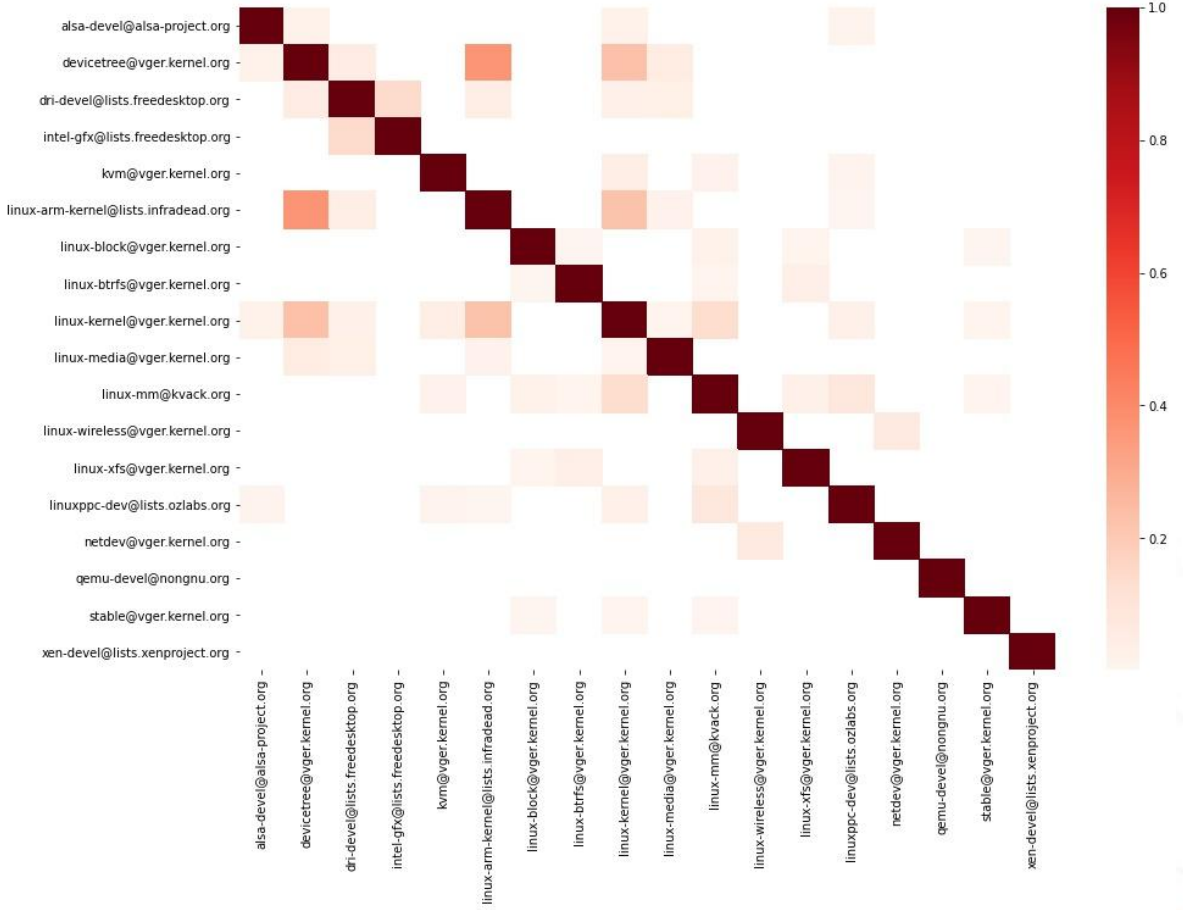
Patches: Individual Mailing Lists

- The positive correlation heatmap of mailing lists are shown in the graph.
- A correlation of 1 would indicate that patches are always sent to the corresponding two mailing lists together



Patches: Individual Mailing Lists

- Correlations of selected mailing lists are shown in the graph
- Some mailing lists are observed to be paired together frequently:
  - linux-arm-kernel@lists.infradead.org and devicetree@vger.kernel.org lists have 0.37 correlation.

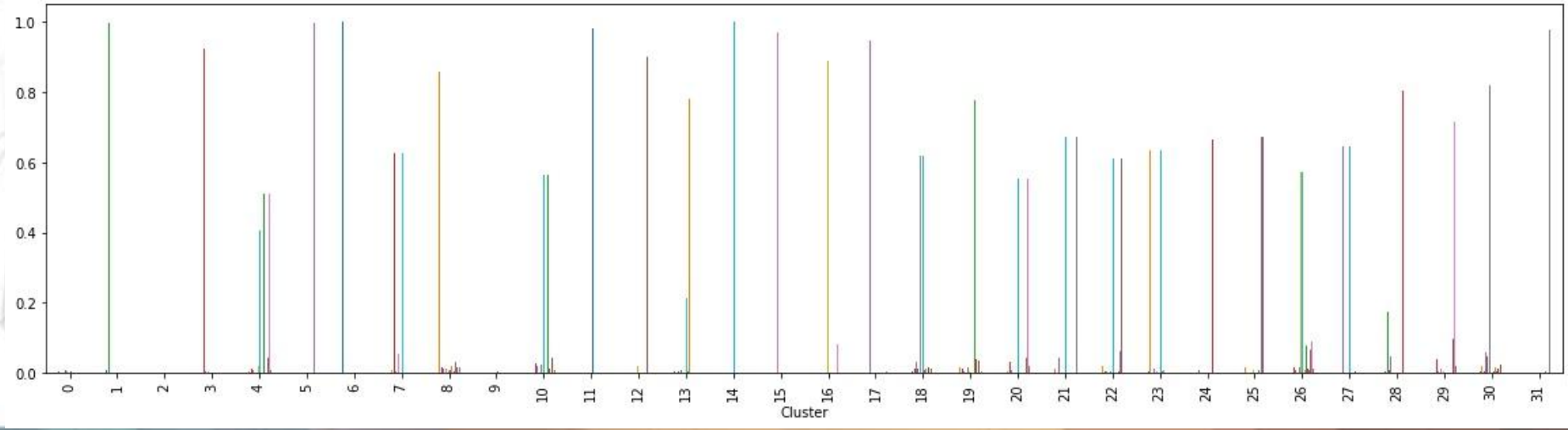


Patches: Individual Mailing Lists

- As patches can be sent to multiple mailing lists, measuring the effects of individual mailing lists on the number of responses becomes a challenge
- New approach: Cluster patches together according to the mailing lists they have been sent to, measure the effects of being in a cluster on the number of responses
- Most defining mailing lists are selected with a variance threshold and used for clustering

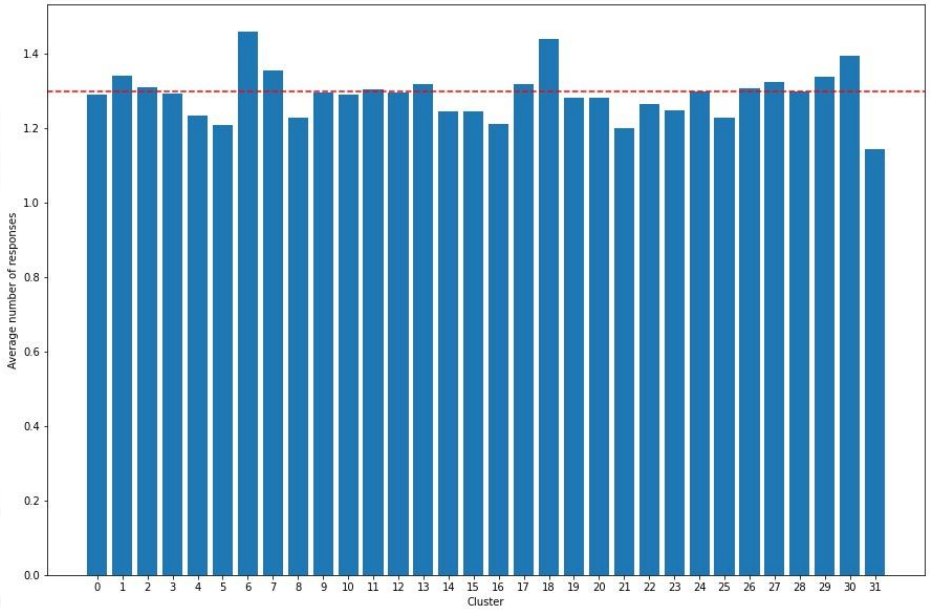
Patches: Individual Mailing Lists

- Using the selected mailing lists, we form 32 clusters of **patches**
- Each cluster has its characteristics in terms of frequent mailing lists patches were sent to

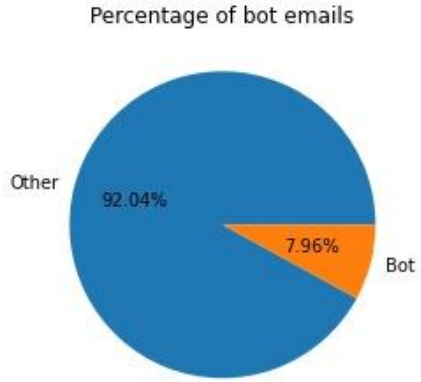


Patches: Individual Mailing Lists

- The graph shows the average number of responses a patch receive for each of the clusters
- The overall average is marked by the red line
- The Kruskal-Wallis test rejects the hypothesis of equal means across clusters with 95% significance

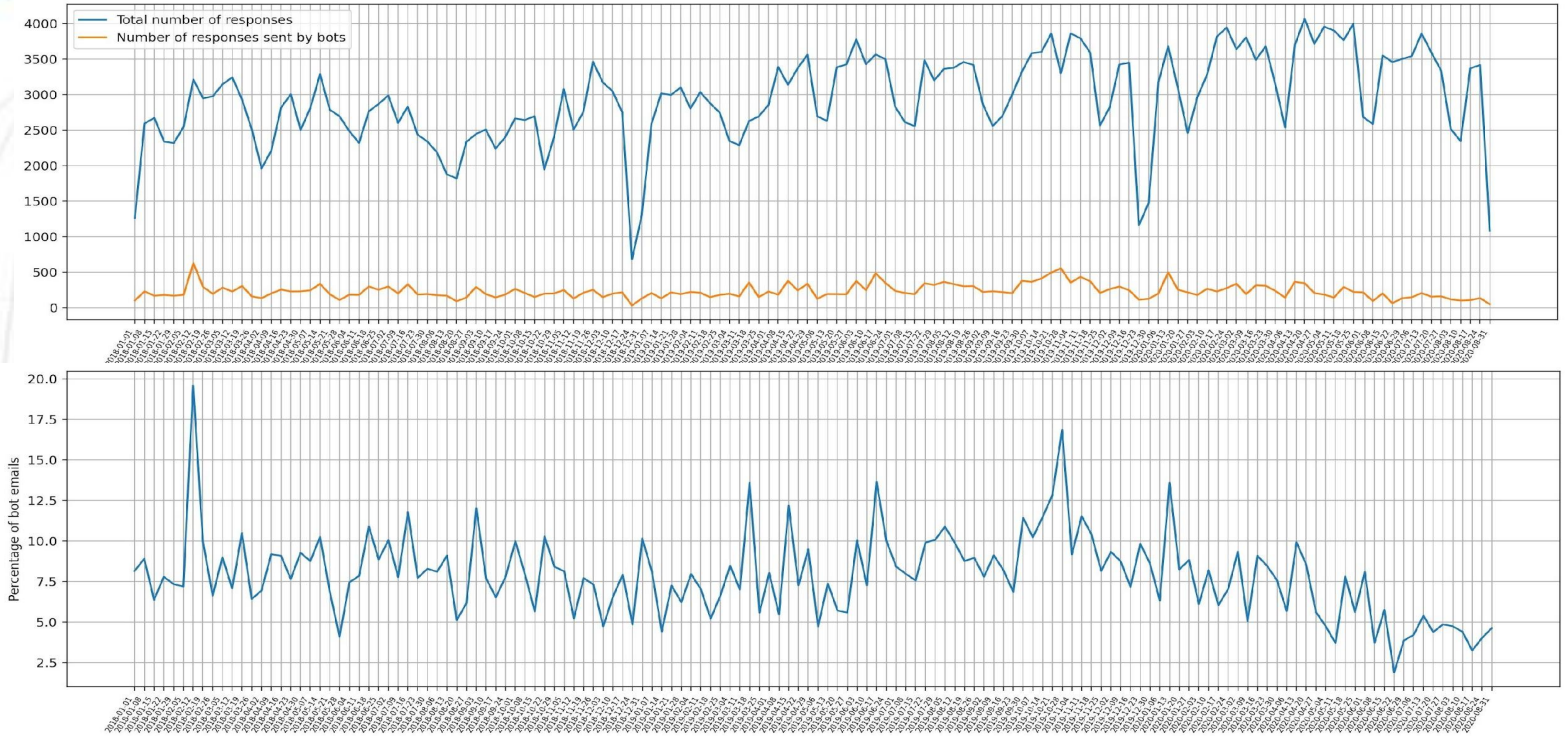


$$H_0 : \mu_i = \mu_j \; \forall i, j \in [0, 31] \qquad H_A : \mu_i \neq \mu_j \; \exists i, j \in [0, 31]$$



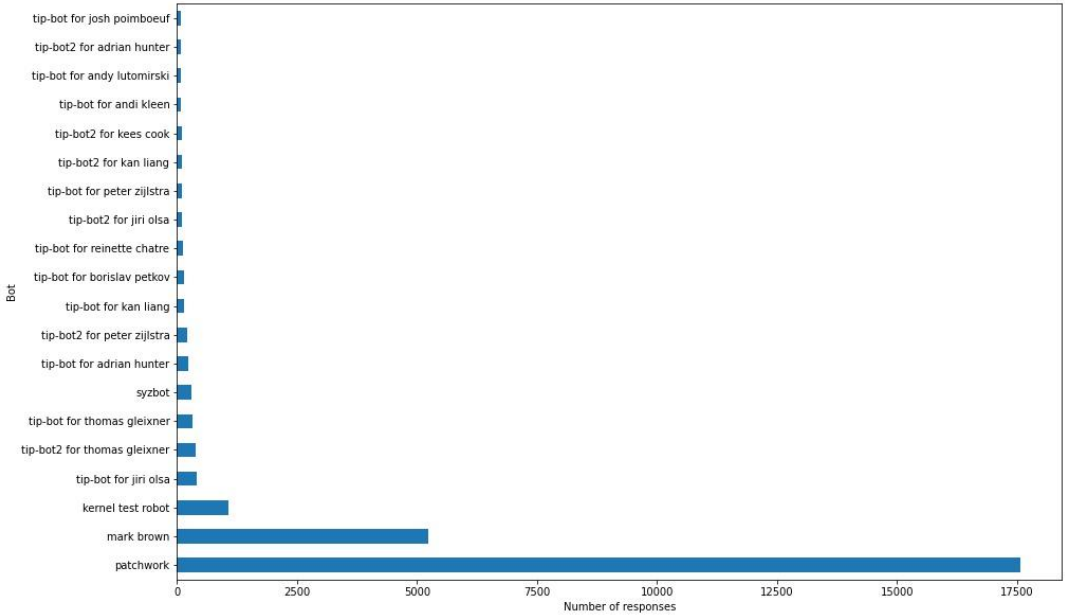
Bots

Bots: Activity Over Time



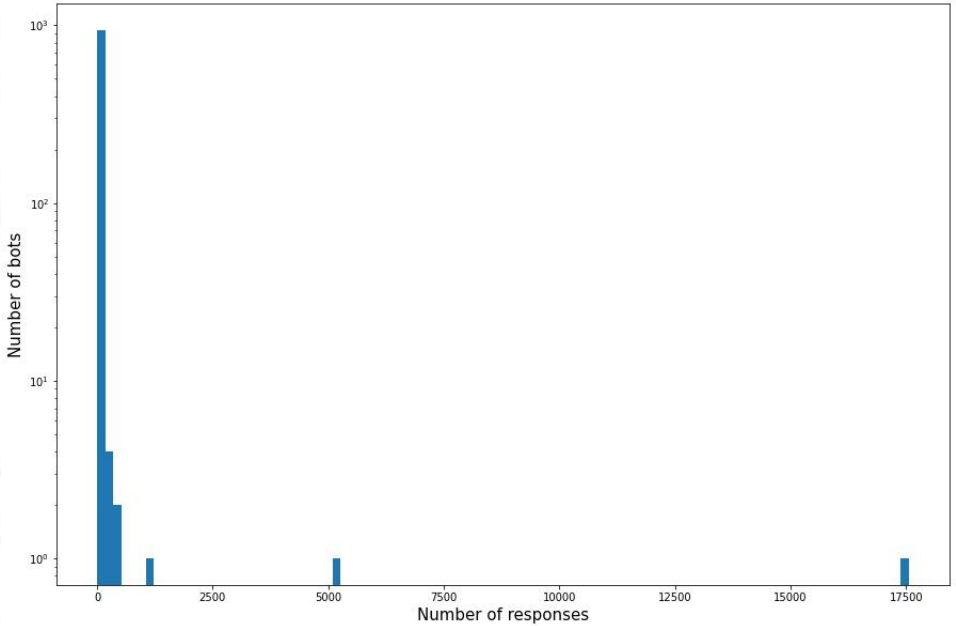
Bots: Most Active Bots

- The patchwork bot has sent all emails to intel-gfx@lists.freedesktop.org
- The bot for Mark Brown is significantly active on alsa-devel@alsa-project.org and linux-kernel@vger.kernel.org lists.



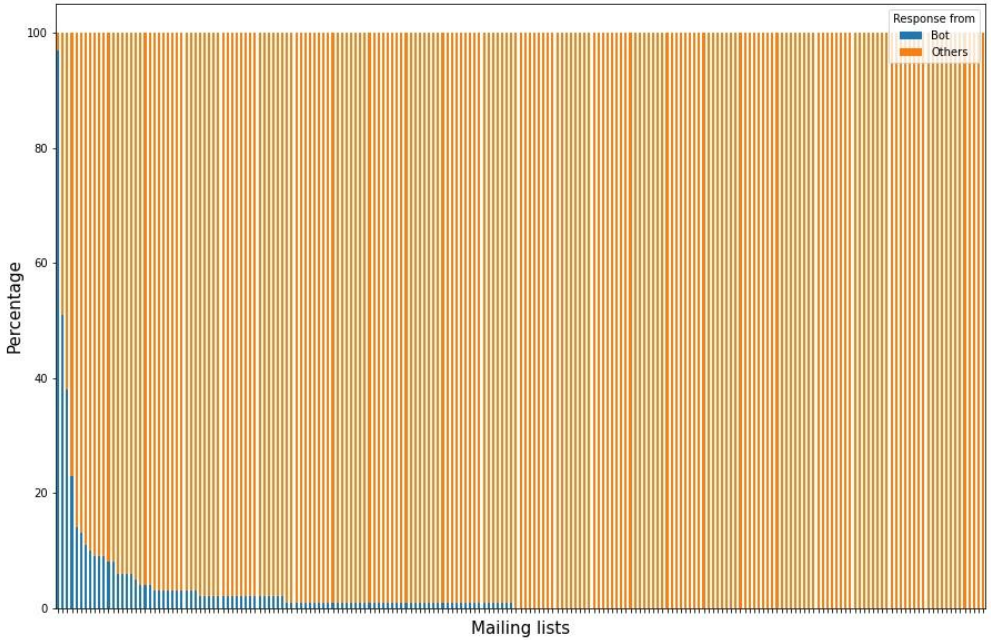
Bots: Distribution of Number of Responses per Bot

- Other than exceptions such as the Patchwork bot and the bot for Mark Brown individual bots have sent very few emails.
- The amount of bots existing in a mailing lists may not correspond to increased bot activity in the mailing list.



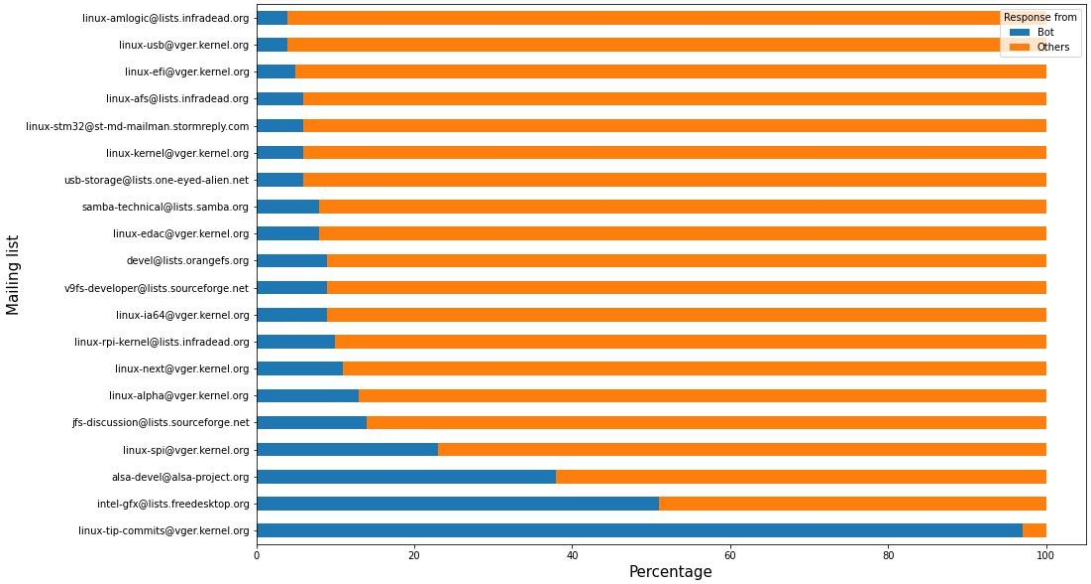
Bots: Activity Across Mailing Lists

- The graph shows the percentages of emails sent by bots across different mailing lists.
- Only the top three of the mailing lists has larger than 30% of their email activity coming from bots.



Bots: Activity Across Mailing Lists

- The graph shows the mailing list with the highest percentage of bot emails.
- 11.94% of all of the emails sent to kernel development community were sent to the top three mailing lists.
- linux-tip-commits@vger.kernel.org has only 0.83% of all of the emails.



LINUX PLUMBERS CONFERENCE September 20-24, 2021

Thank You!

Resources

- Submitting patches: the essential guide to getting your code into the kernel, 2021. url: https://www.kernel.org/doc/html/latest/process/submitting-patches.html (visited on 02/23/2021).
- How the development process works, 2021, url: https://www.kernel.org/doc/html/latest/process/2.Process.html#the-lifecycle-of-a-patch (visited on 03/08/2021).
- J. Corbet. Gitdm (the "git data miner"), git://git.lwn.net/gitdm.git, 2020.
- J. Corbet. Gitdm (the "git data miner"), https://lwn.net/Articles/290957/, 2008.
- HOWTO do Linux kernel development, 2021. url: https://www.kernel.org/doc/html/latest/process/howto.html (visited on 02/23/2021).
- C. M. Bishop. Pattern Recognition and Machine Learning, Springer-Verlag New York, 2016.
- T. Lumley. Biostatistics: a methodology for the health sciences, Wiley-Interscience, 2004.

Backup

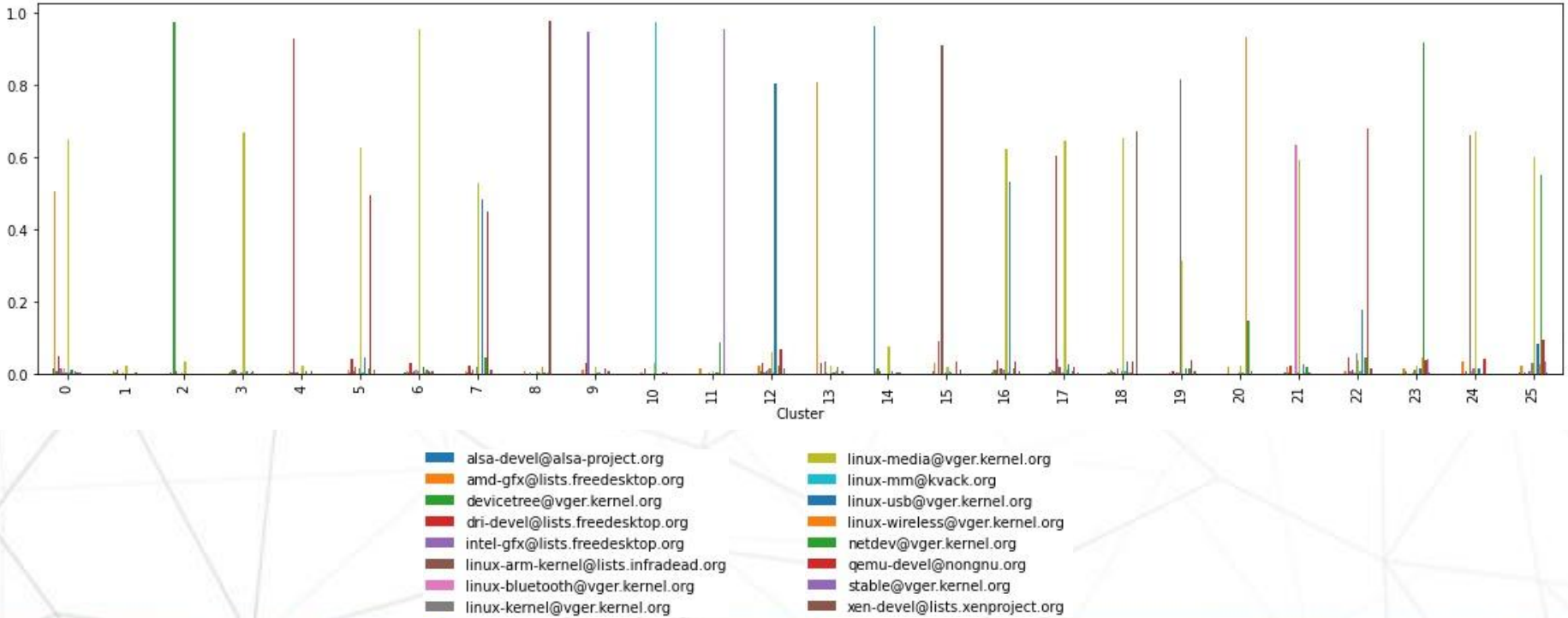
Activity Area

Authoring Activity: Activity Area

- Each person has a vector of mailing lists. Each dimension value is the number of emails sent to the respective list.
- Vector is normalized to have the norm of 1.
- More defining mailing lists are selected to be used for clustering people using a variance threshold.

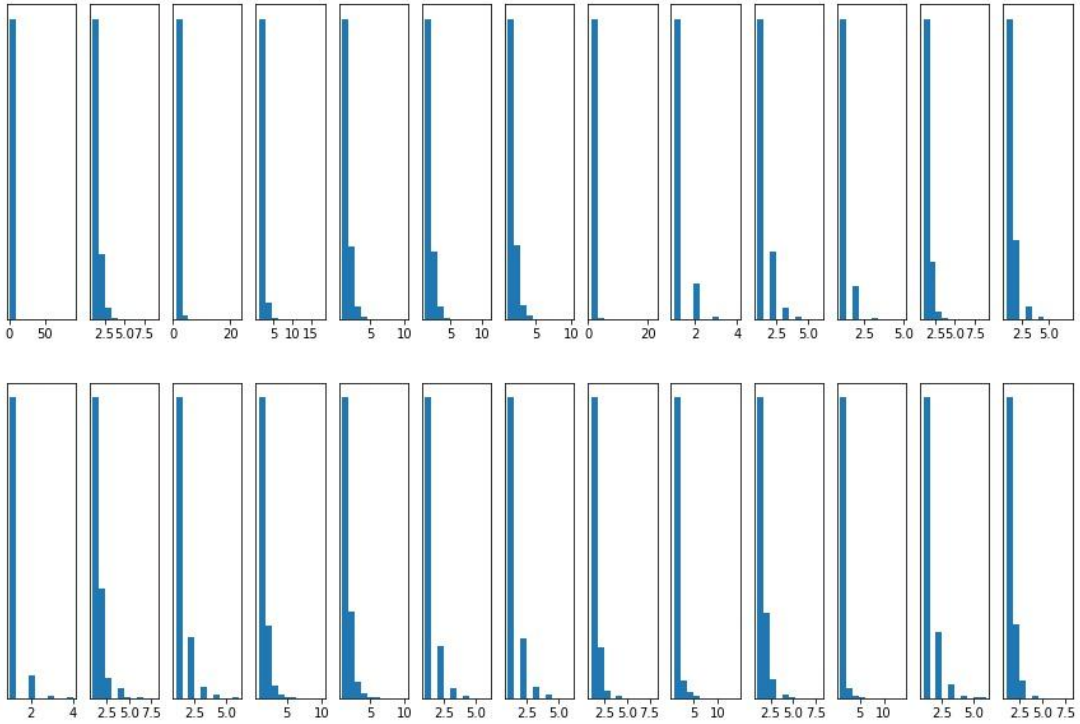
Id	alsa-devel@alsa-project.org	amd-gfx@lists.freedesktop.org	devicetree@vger.kernel.org	dri-devel@lists.freedesktop.org	intel-gfx@lists.freedesktop.org	linux-arm-kernel@lists.infradead.org	linux-kernel@vger.kernel.org	linux-media@vger.kernel.org	linux-mm@kvack.org	linux-tip-commits@vger.kernel.org	linux-usb@vger.kernel.org
20	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	0.000000	0.0	0.0	0.00000
21	0.0	0.0	0.000000	0.0	0.0	0.000000	0.577350	0.000000	0.0	0.0	0.00000
22	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	1.000000	0.0	0.0	0.00000
23	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	0.000000	0.0	0.0	0.00000
24	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	0.000000	0.0	0.0	0.00000
25	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	0.000000	0.0	0.0	0.00000
26	0.0	0.0	0.000000	0.0	0.0	0.000000	0.000000	0.707107	0.0	0.0	0.00000
27	0.0	0.0	0.577350	0.0	0.0	0.577350	0.000000	0.000000	0.0	0.0	0.57735
28	0.0	0.0	0.000000	0.0	0.0	0.000000	0.577350	0.000000	0.0	0.0	0.57735
29	0.0	0.0	0.258199	0.0	0.0	0.258199	0.774597	0.258199	0.0	0.0	0.00000

Authoring Activity: Activity Area



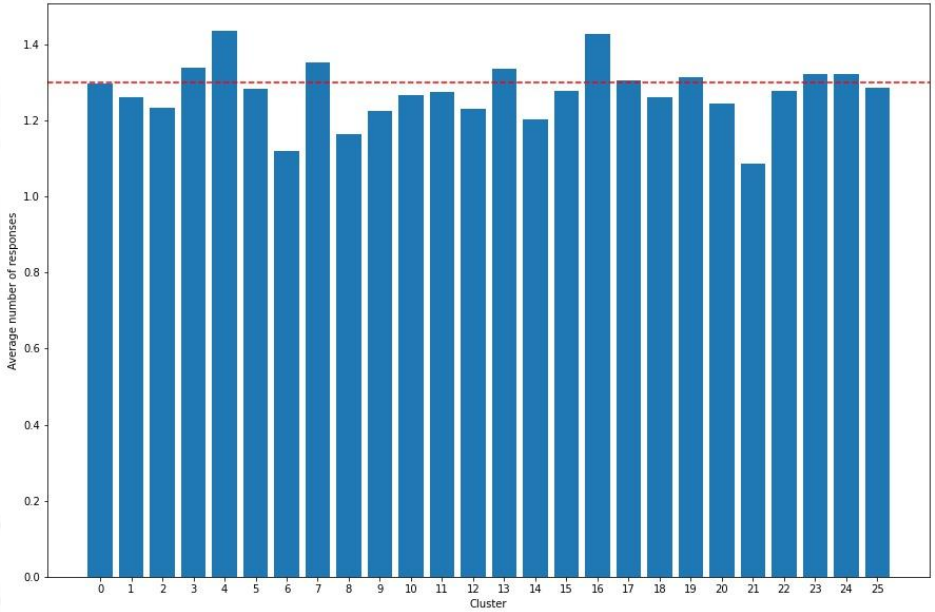
Authoring Activity: Activity Area

- The graphs show the distributions of the number of responses per patch authored by developers in each clusters
- Since the distributions are not normal, a non-parametric test is conducted for differences of means across each cluster



Authoring Activity: Activity Area

- Kruskal-Wallis test rejects the hypothesis of equal means across clusters and concludes that there is at least two groups with equal means
- We disprove that the author's area of activity does not make a difference



$$H_0 : \mu_i = \mu_j \; \forall i, j \in [0, 25] \quad H_A : \mu_i \neq \mu_j \; \exists i, j \in [0, 25]$$

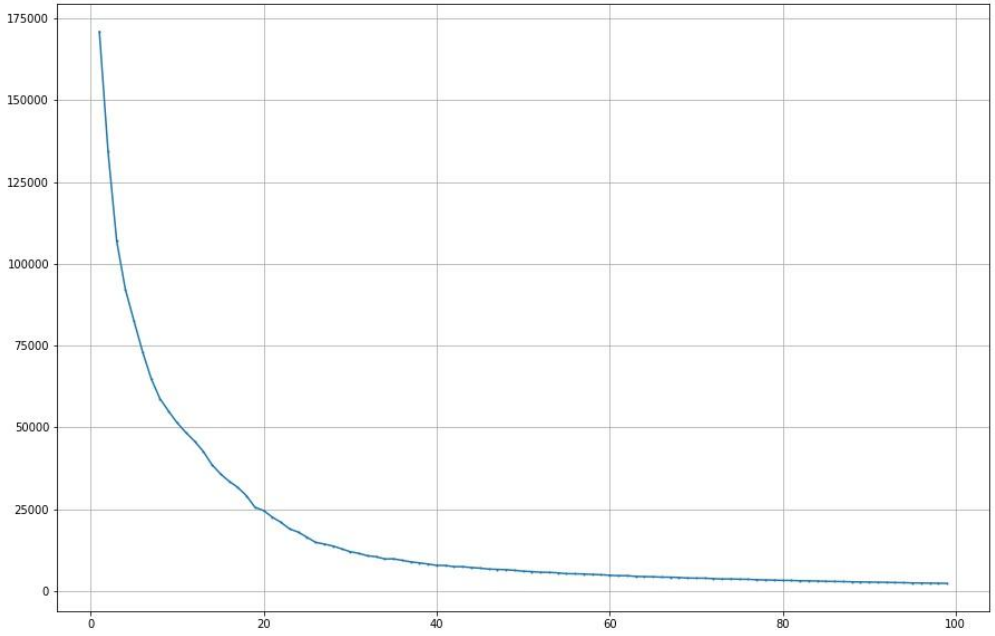
Backup: Patch Vectors for K-Means and Correlations

- Each patch has a vector of 1s and 0s according to which mailing lists they have been sent to, dimensions are the mailing lists
- Some dimensions are removed with a variance threshold

	alsa-devel@alsa-project.org	amd-gfx@lists.freedesktop.org	apparmor@lists.ubuntu.com	ath10k@lists.infradead.org	autofs@vger.kernel.org	b.a.t.m.a.n@lists.open-mesh.org	b43-dev@lists.infradead.org	bpf@vger.kernel.org	bridge@lists.linux-foundation.org	cake@lists.bufferbloat.net	ccan@lists.ozlabs.org
patch_id											
<2X21xsgg8bEyTVVoBik7GT2JM5chtujJdSSpMBd-J7xIRMuRPqnF8Or25Al9oNj5R08x57BAEhkOUdT1-8FEmtoGVKXOB7X1DTILztKep-8=@krauser.org>	0	0	0	0	0	0	0	0	0	0	0
<-LYZxtmyBTf36wklyxa0PphDQ1FecAgEF7TMnSvyCm9Y_EFmz-4AUROX6qc4HKUjomE0HumDgVrSIbHsUMJnRSrBR2c3gPCVDNUmz7klPkE=@emersion.fr>	0	0	0	0	0	0	0	0	0	0	0
<0-v1-4eb72686de3c+5062-hmm_no_flags_jgg@mellanox.com>	0	1	0	0	0	0	0	0	0	0	0
<0-v1-982a13cc5c6d+501ae-fix_best_pgsz_stub_jgg@nvidia.com>	0	0	0	0	0	0	0	0	0	0	0

Patches: Individual Mailing Lists

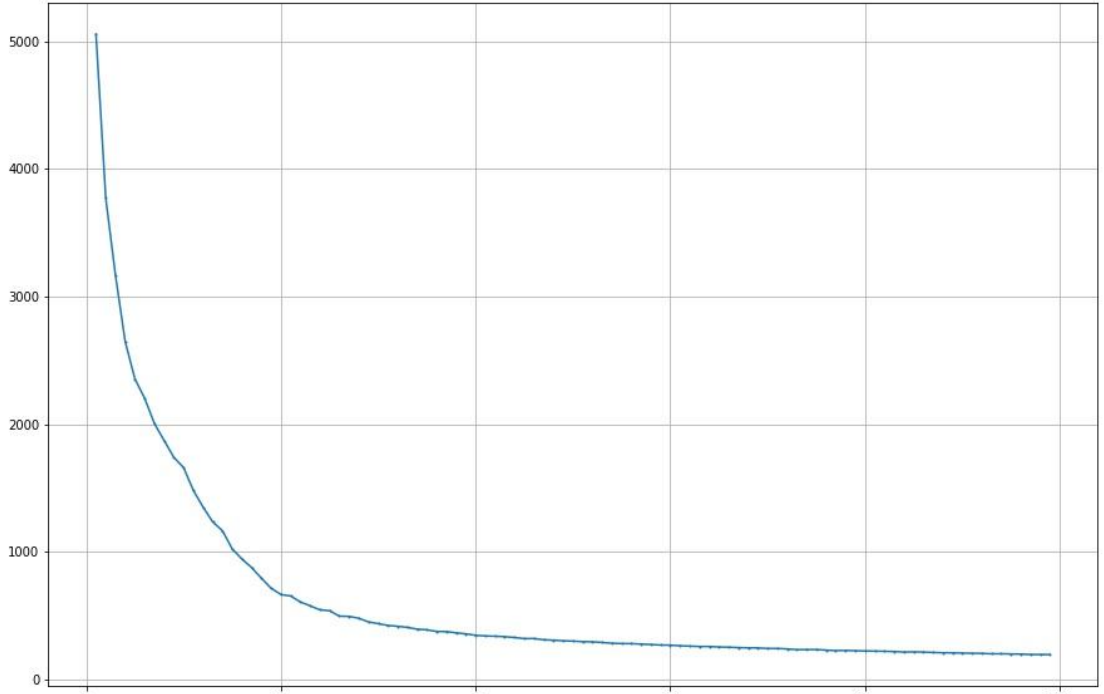
- Using the same sum of distances criteria as the previous clustering, the number of clusters is selected to be 32



Errors with increased number of clusters

Authoring Activity: Activity Area

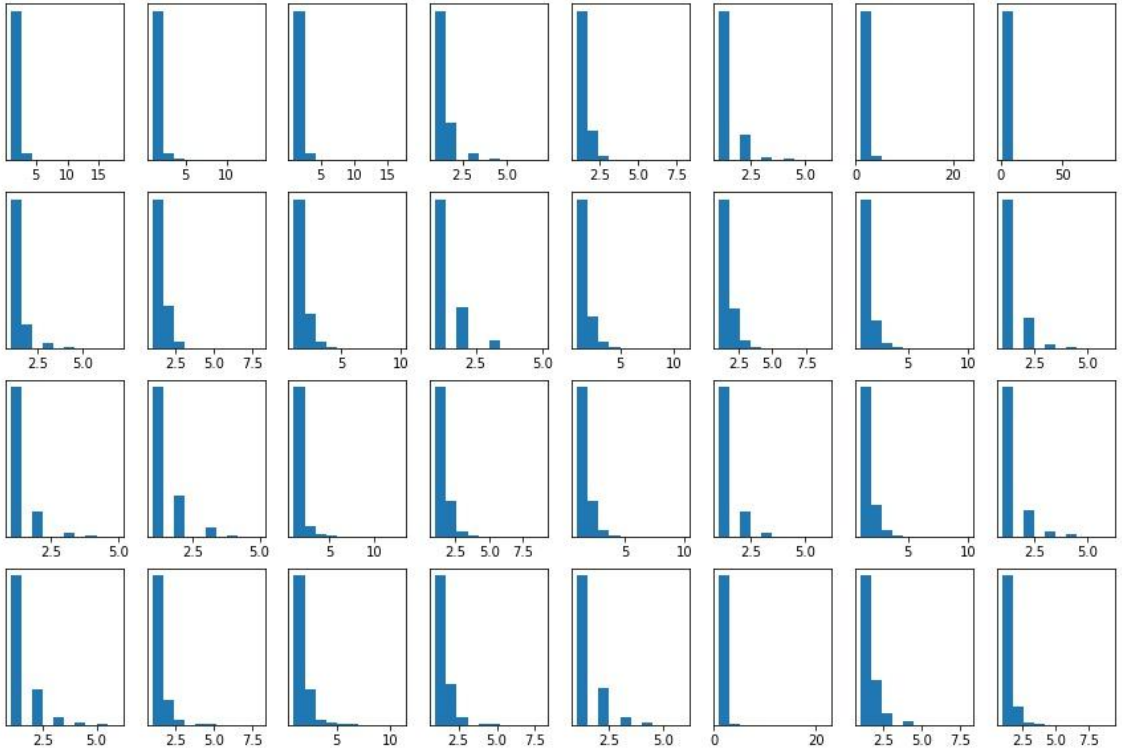
- For clustering, K-means clustering algorithm is used.
- The algorithm works by recursively assigning data points to cluster centers and calculating new cluster centers
- The algorithm requires to specify the number of clusters.
- Error measure is the sum of distances from data points to corresponding centers.



Errors with increased number of clusters

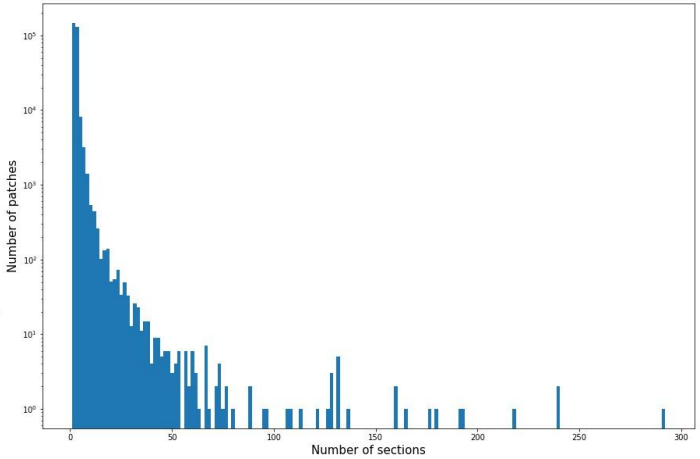
Patches: Individual Mailing Lists

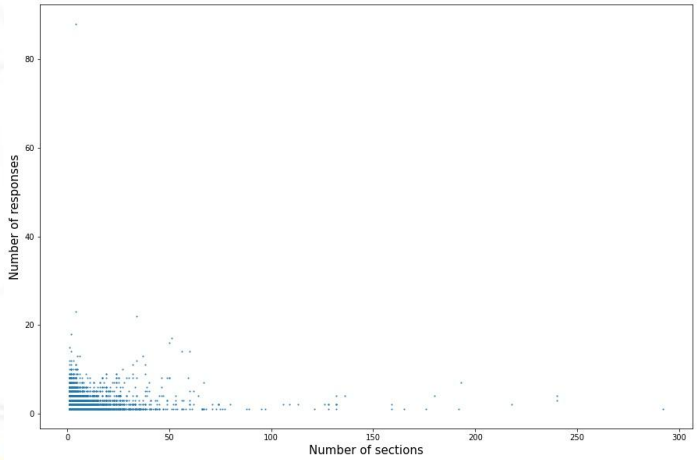
- The distributions of the number of responses per patch is shown in the graph
- Since the distributions are not normal, a non-parametric test is conducted for differences of means across each cluster



Patches: Number of Sections

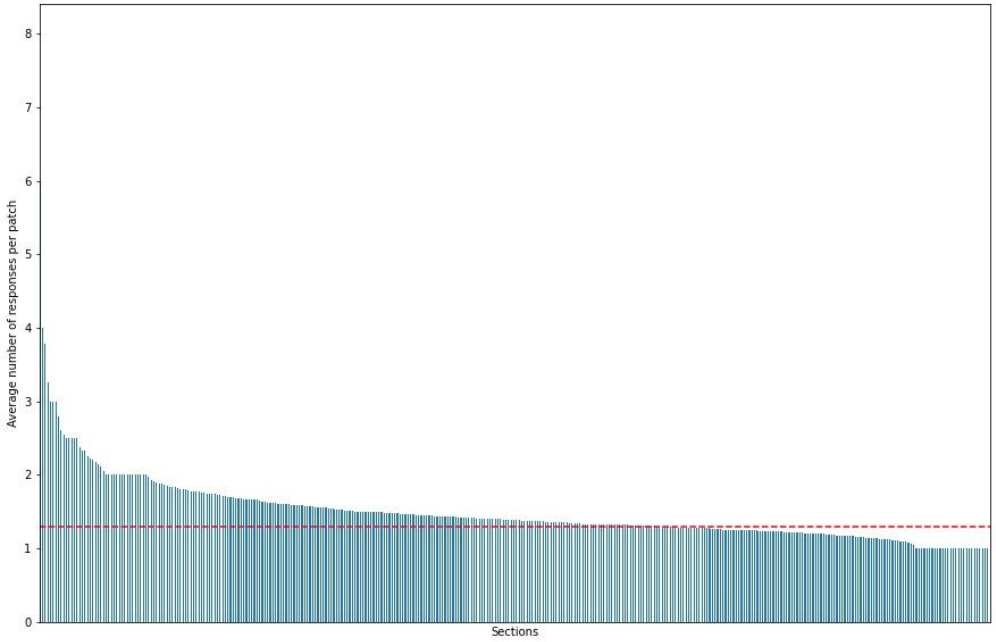
- The log-scaled histogram shows the distribution of number of sections related to a patch, many has fewer sections.
- In the graph, many patches are clustered in the left side, while no relation to number of responses is observed.





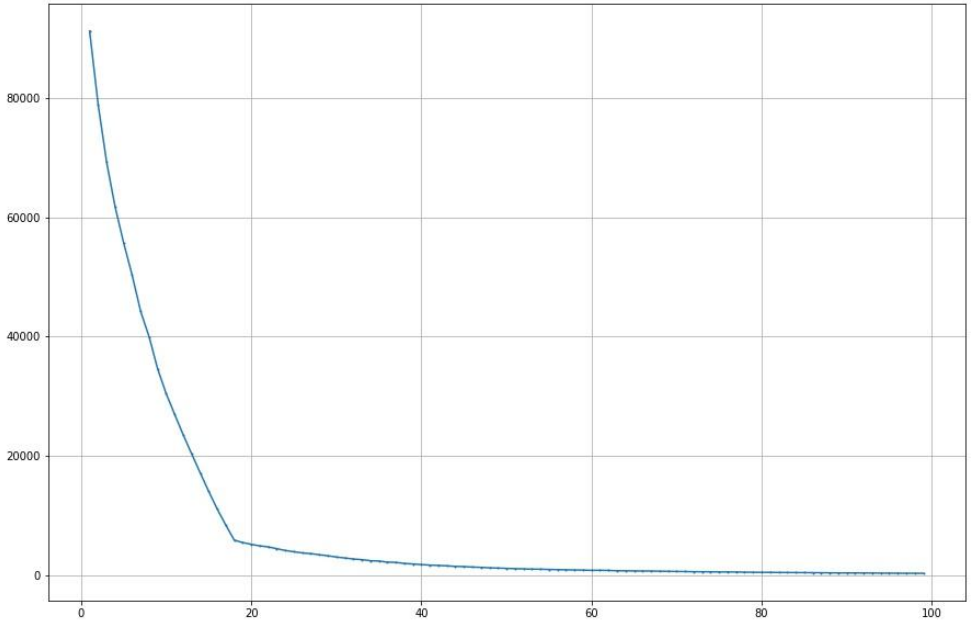
Patches: Individual Sections

- The graph shows the average number of responses per patch for each of the maintainer sections.



Patches: Individual Sections

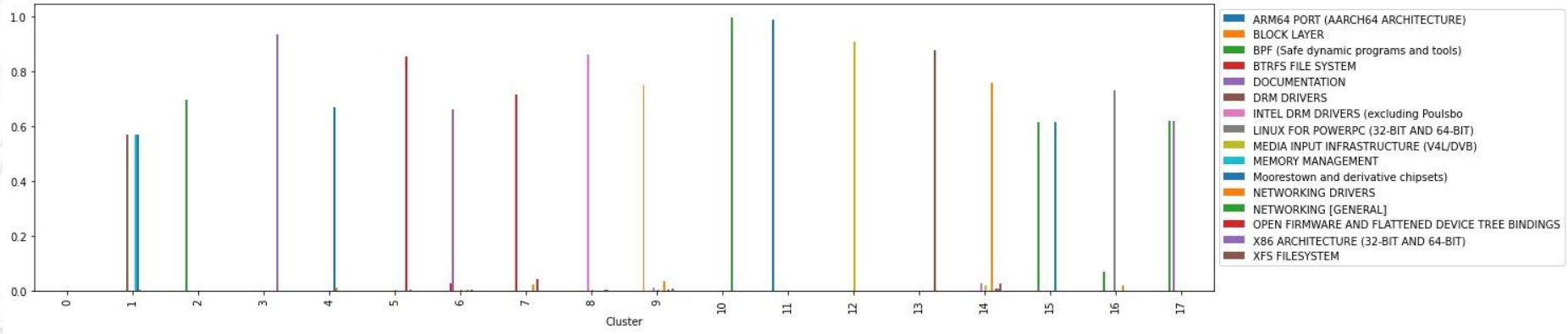
- Using the same sum of distances criteria as the previous clustering, the number of clusters is selected to be 18



Errors with increased number of cluster

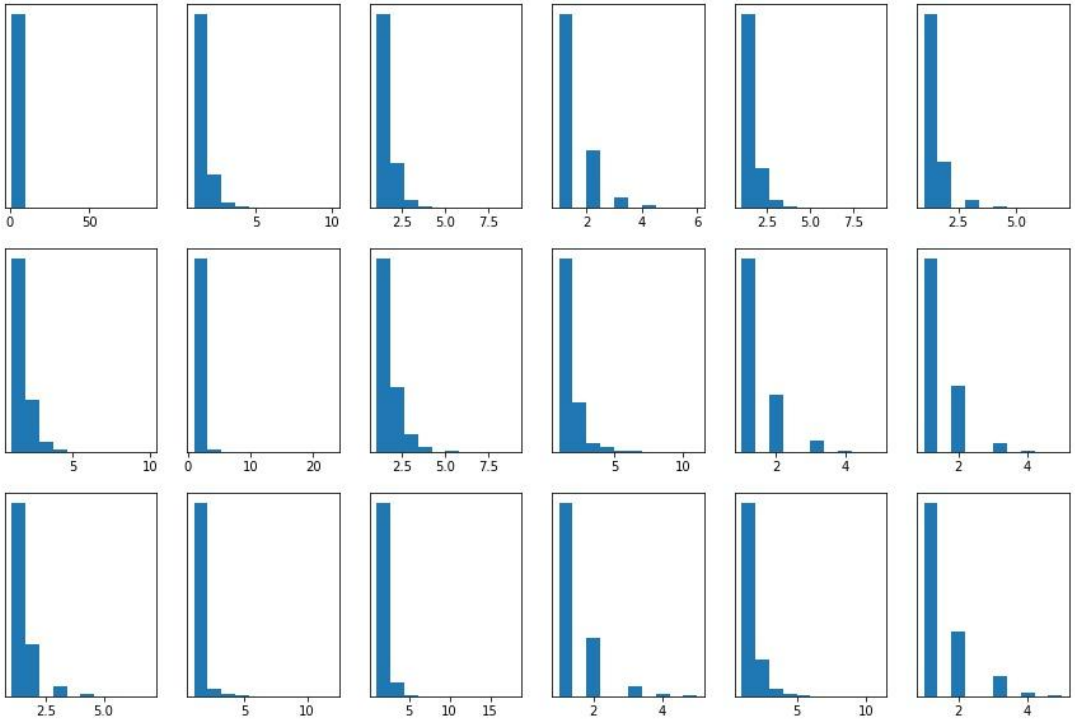
Patches: Individual Sections

- Coefficients of the center vectors of each cluster is shown below.



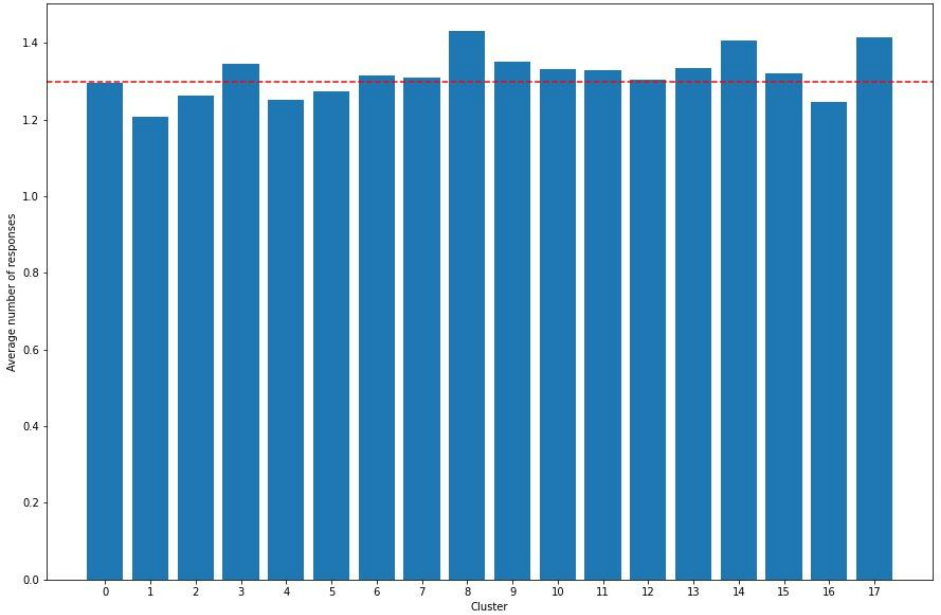
Patches: Individual Sections

- Similar to mailing list case, the distributions of the number of responses per patch within clusters are not normal
- Non-parametric Kruskal-Wallis test is conducted



Patches: Individual Sections

- The graph shows the average number of responses a patch receive for each of the clusters
- The overall average is marked by the red line
- Similar to the mailing lists case, the Kruskal-Wallis test rejects the hypothesis of equal means across clusters with 95% significance



$$H_0 : \mu_i = \mu_j \; \forall i, j \in [0, 17] \qquad H_A : \mu_i \neq \mu_j \; \exists i, j \in [0, 17]$$

Backup: Investigating Reversed Quadratic Relationships

- x_1 : the number of mailing lists
- x_2 : the number of files
- Simple formula has an R-Squared value of 0.76
- Extended formula has an R-Squared value of 0.78

$$1.59 * \frac{1}{x_1} + 1.46 * \frac{1}{x_2} - 1.82 * (\frac{1}{x_1} * \frac{1}{x_2}) = y$$

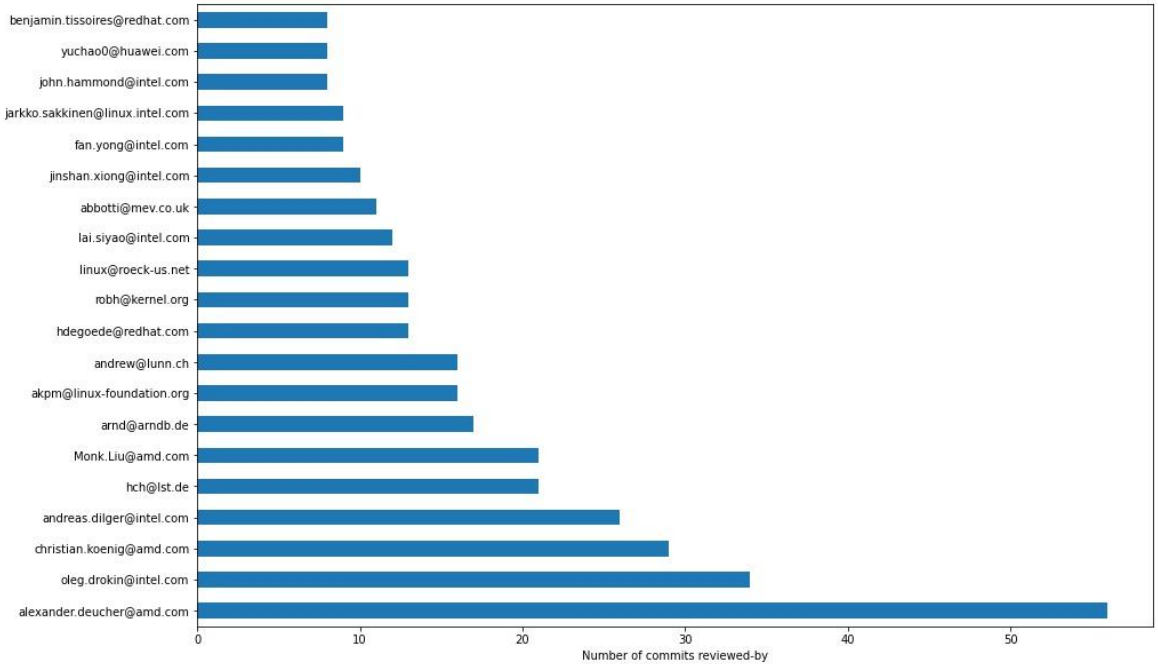
Simple Formula

$$-2.01 * \frac{1}{x_1^2} - 0.8 * \frac{1}{x_2^2} + 3.22 * \frac{1}{x_1} + 1.41 * \frac{1}{x_2} - 0.61 * (\frac{1}{x_1} * \frac{1}{x_2}) = y$$

Extended Formula

Backup: One-Time Committer Reviewers

- The graph shows Reviewed-by tags, most frequently showing up in one time committer commits.



Backup: Average Number of Responses per Patch per Person