Community's Adventures in Analyticsland

Or the State of the Community Through New Analytics

Christelle Zouein & Kevin Ottens

Akademy 2021, June 25th

HAUTE COUTURE

enioka

whoami

- Computer engineering student
- Majoring in data science
- Internship at enioka Haute Couture
  - Under Kevin Ottens
  - Working on ComDaAn in 2019
- Data science consultant intern at enioka consulting
- Based in Paris

whoami

- Started to use KDE with 1.0-beta1 in 1997
- Procrastinated until 2003 to finally contribute code
- Fell in love with the community back then
- Kept doing things here and there... most notably helped with:
  - kdelibs
  - KDE Frameworks architecture
  - the KDE Manifesto
  - Community Data Analytics
- Part of the enioka Haute Couture family
- Living in Toulouse

Let's Start Our Journey

Previously

Looking Humongous
Looking At Communities

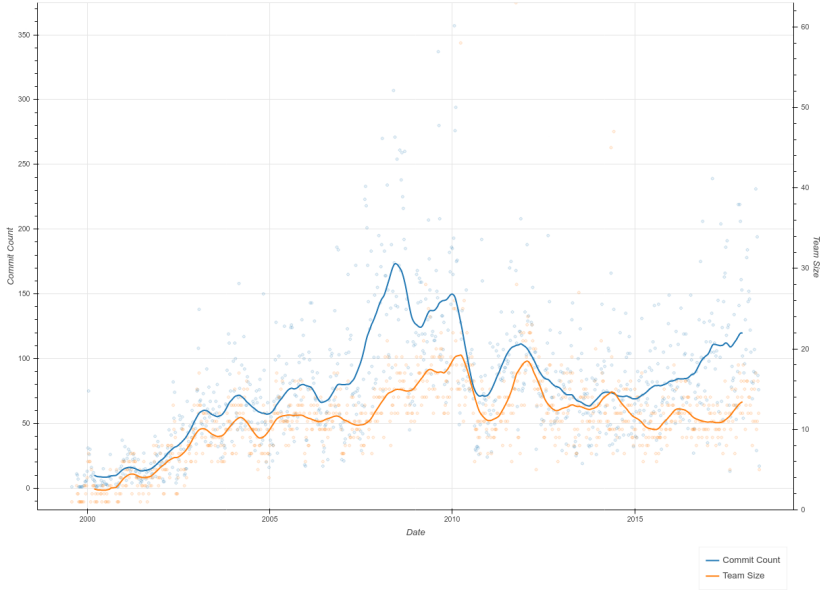
- Took over the “green blobs” idea from Adriaan de Groot
- Author of git-viz, and made the blobs *blue*
- Delivered interesting talks on the topic, go watch them
- Retired from KDE in 2017
- Now having fun with his humongous dog in Berlin

Introducing ComDaAn

- Forked from git-viz
- Reused mostly the git parsing code
- Otherwise swapped most of the dependencies
  - Pandas for the data processing
  - Networkx for the graph analyses
  - Bokeh for the output
- Added ways to clean up the data with rules
- Introduced new visualizations
- More interactivity
- Easier to explore the results

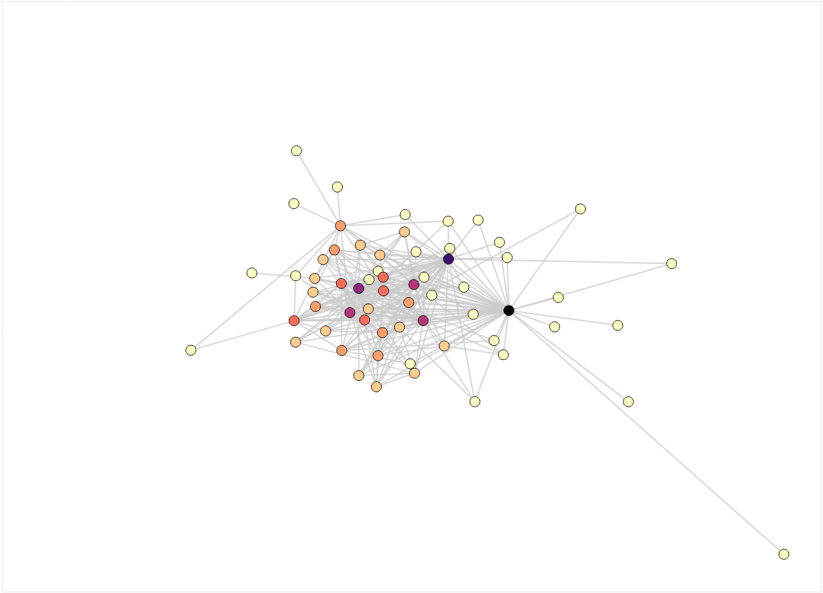
Let's see a few examples...

Team Size Plot Example

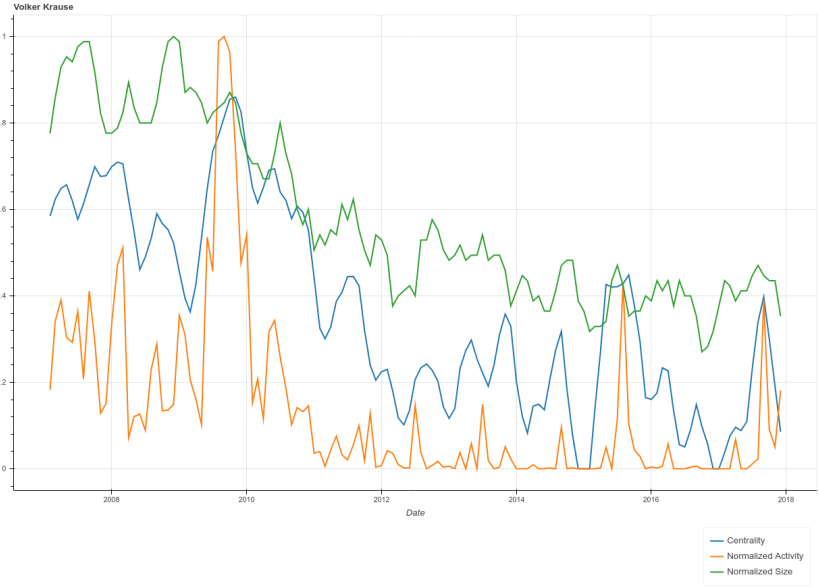


Activity Plot Example

Contributor Network Example



Contributor Centrality Example



Just a Fun Puzzle?

- Definitely a fun puzzle... but not only!
- Can show how healthy a community is
  - Activity level
  - Recruiting contributors
  - Contributor retention
  - Bus factor
  - Team structure
  - Team splits
- And also professional uses...

Professional Uses

- Some of them are side effects of the community health checks
- Project technical framing
  - Quite a bit about selecting building blocks and dependencies
  - Need to maximize chances to pick something durable
  - Community health over time can be a good indicator
- Code auditing
  - Easier to explore project history
  - Evaluate developers turn over
  - Find out how the team is structured
    - Who owns what
    - Who works with whom
  - Identify project key personnel
    - Produce project specific “who’s who”

Following The Rabbit

What's new

The Rabb^WElephant in The Room

- It's only about the code commits...
- Obviously not capturing lots of a project life
- New data sources needed
- Enters Christelle...
- An internship to fix some of it

Move Away From Executables

- ComDaAn was a bunch of Python scripts
- Now it is a Python API
  - Better for reproducibility
  - High level API to express intent about what to visualize

```
import comdaan as cd

data = cd.parse_repositories("~/Repositories")
a = cd.activity(data, "id", "author_name", "date")
cd.display(a, output="activity.html", title="Activity All Time")
```

More Data Sources

- Mailing lists
  - Although it's surprisingly cumbersome to get archives in an exploitable format
  - Unlikely to be used much in practice
- GitLab Discussions
  - Same format for both Issues and Merge Requests
  - Means we support both out of the box
  - Companion script provided to pull the data
  - Leads to a new visualisation: responsiveness

- Potentially will allow aggregated views
  - Contributor network if we look at both commits and MRs at the same time?
  - Difficulties to remap commit authors to GitLab users though

Now that KDE transitioned to GitLab, this gets interesting...

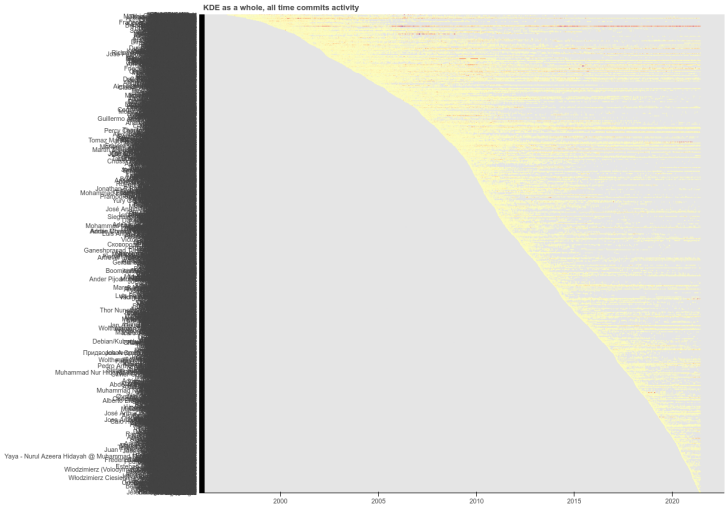
Let's Widen Our Horizons

A Word About Our Dataset

- All of the KDE repositories
  - All 950+ of them
  - Yes, even the ones in unmaintained!
- Rules to cleanup authors identity in commits
- Rules to exclude robots from commits
- Rules probably not exhaustive
  - Try to address the "bigger offenders"
  - Reduces stats biases

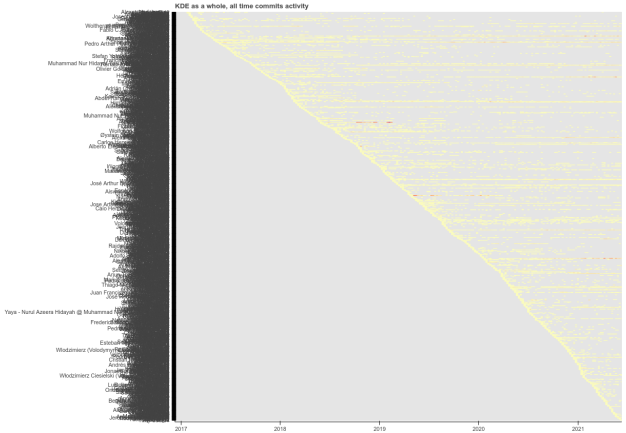
An Update About KDE as a Whole

All Time Activity (Commits Only)



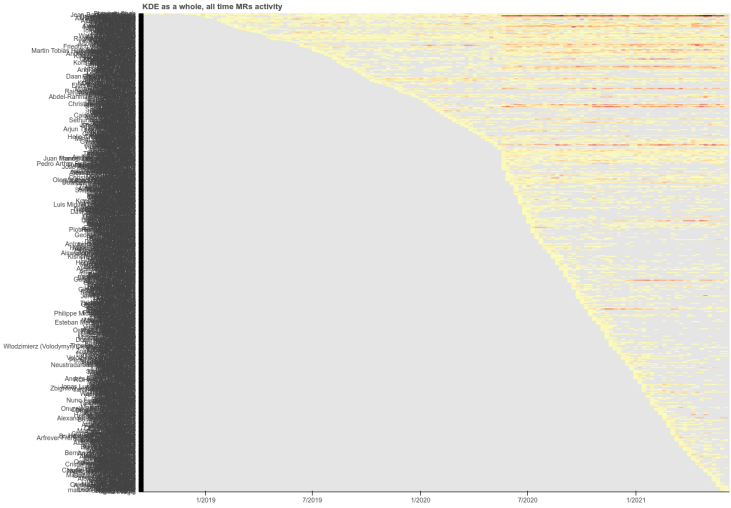
- Laurent Montel, 1999
- Recruiting
  - Faster after 2010
  - Even faster in the last couple of years
- Retention
  - Not great after 2010
  - Much better in the last couple of years

2017 and Following Main Recruits (Commits Only)



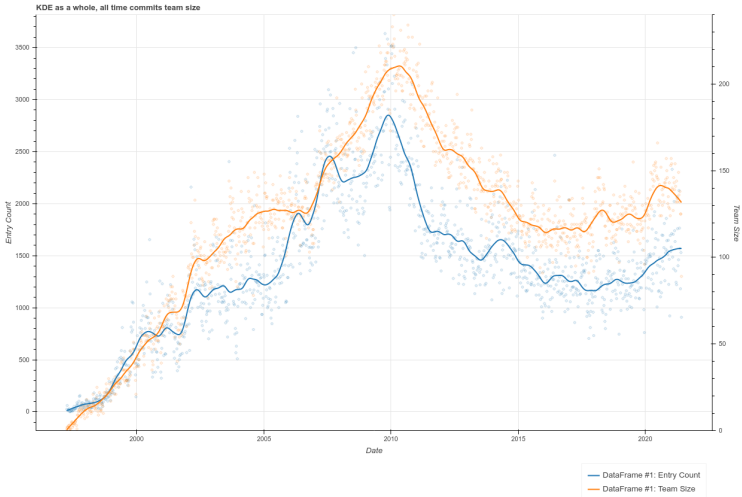
- People we see appear
  - Agata Cacko
  - Ahmad Samir
  - Alexander Lohnau
  - Alexander Stippich
  - Camilo Higuita
  - Carl Schawn
  - David Redondo
  - Devin Lin
  - Han Young
  - Jan Blackquill
  - Jonah Brüchert
  - Méven Car
  - Nate Graham
  - Nicolas Fella
  - Noah Davis
  - Sharaf Zaman
  - Vlad Zahorodnii
  - Waqar Ahmed

All Time Activity (GitLab MRs)



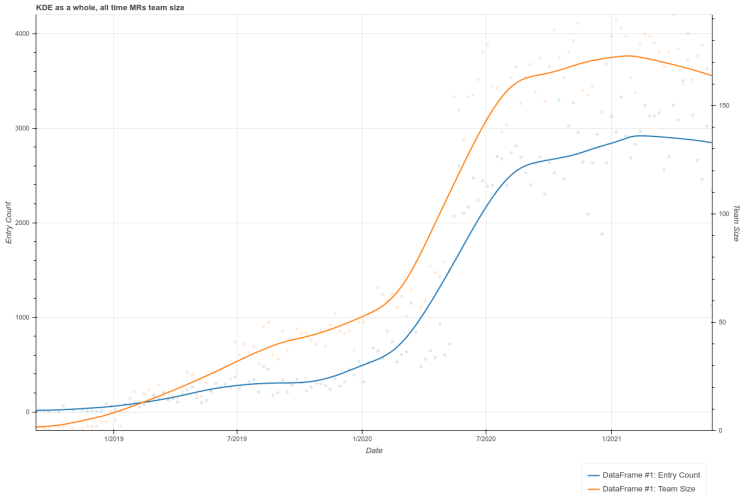
- Less history obviously
- Clearly we see GitLab picking up
- More people
  - Xaver Hugl
  - Mikel Johnson
- We didn't spot them earlier, why?

All Time Team Size (Commit Only)



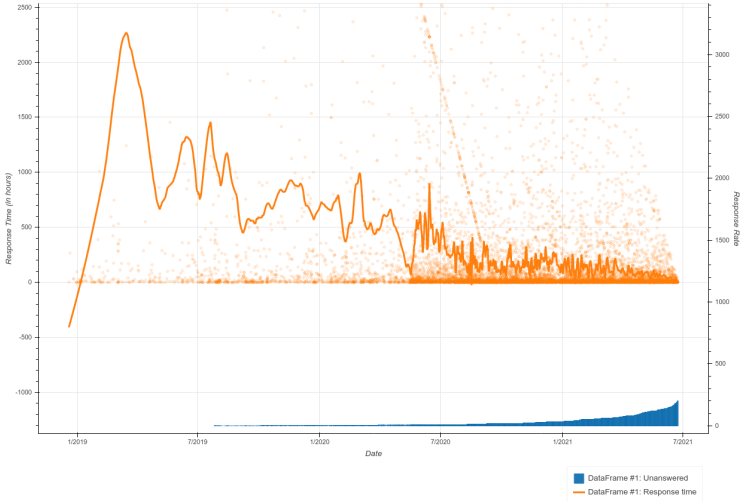
- The Nokia peak still visible around 2010
- Then we see it stabilized around 2016
- And clearly it's picking up again since 2019/2020

All Time Team Size (GitLab MRs)



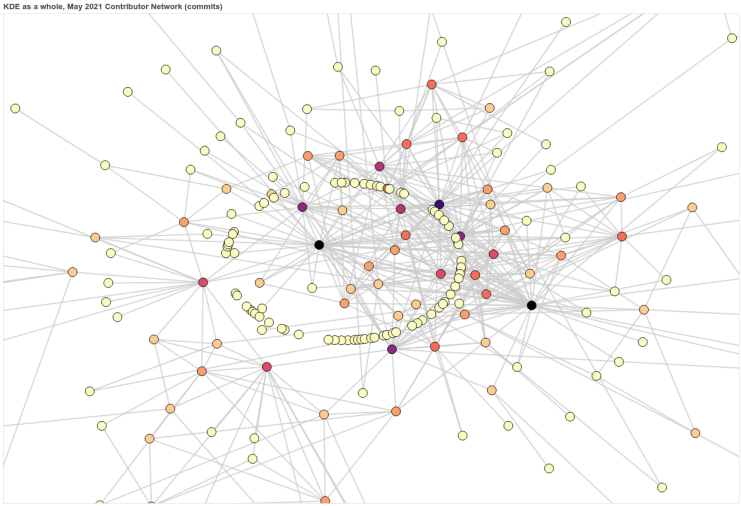
- We see it being picked up during 2020...
- Need to accumulate data still

All Time Responsiveness (GitLab MRs)



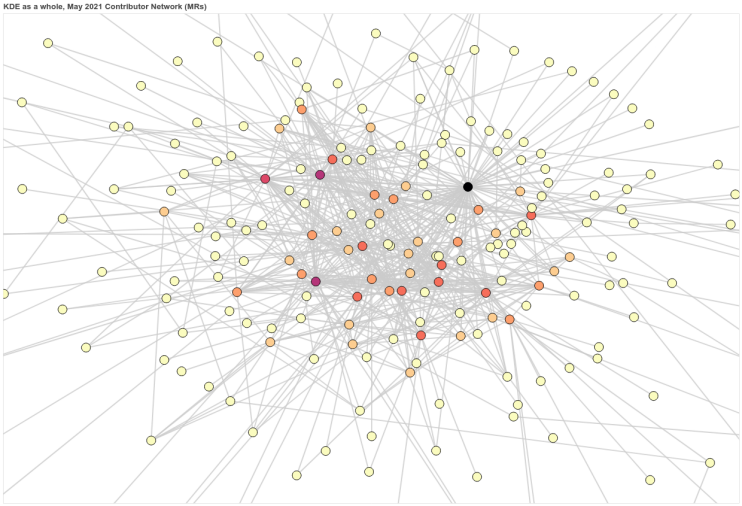
- Took some time to get in people's habits
- Converges to being somewhat responsive (around a couple of days)
- Stock of unanswered MRs tends to build up
- Anecdote: history goes before the instance existed
  - Kaidan got imported...
  - Has been excluded to produce that plot

May Network (Commits Only)



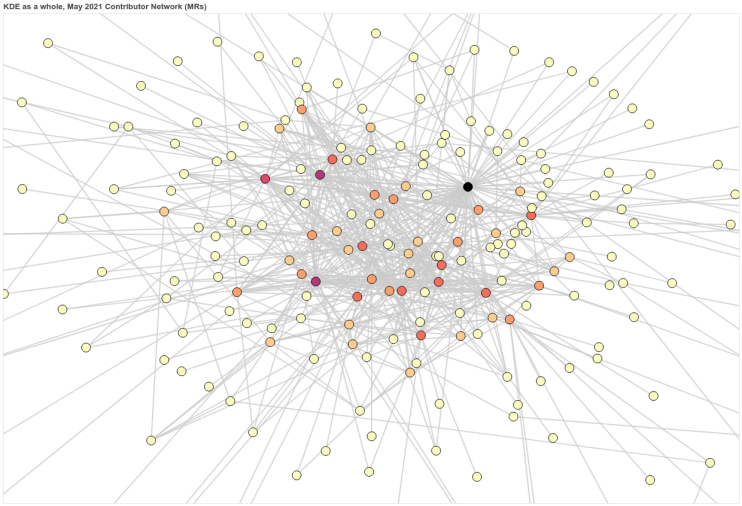
- Centrality top 5
  - Laurent Montel
  - Alexander Lohnau
  - Nicolas Fella
  - Carl Schawn
  - Heiko Becker

May Network (GitLab MRs)



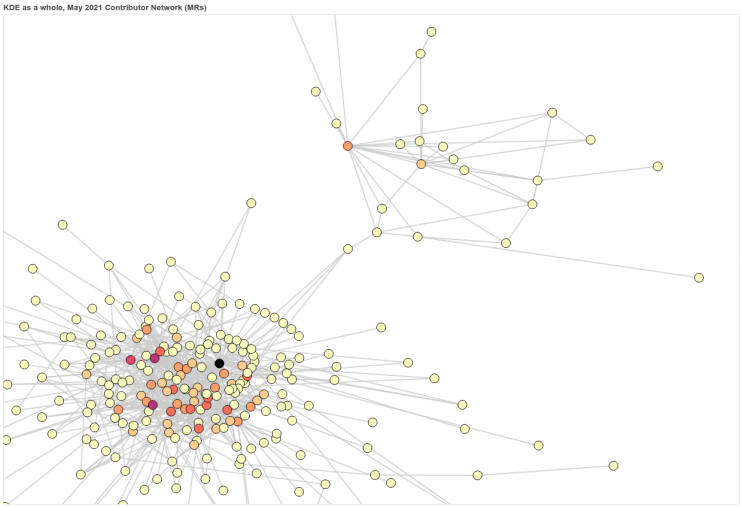
- Centrality top 5
  - Nate Graham
  - Nicolas Fella
  - Alexander Lonhau
  - Ahmad Samir
  - Aleix Pol
- Laurent is nowhere to be seen
  - Shows the "commits in the same file" assumption breaks down
- Let's zoom out a bit...

May Network (GitLab MRs)



- Centrality top 5
  - Nate Graham
  - Nicolas Fella
  - Alexander Lonhau
  - Ahmad Samir
  - Aleix Pol
- Laurent is nowhere to be seen
  - Shows the "commits in the same file" assumption breaks down
- Let's zoom out a bit...

May Network (GitLab MRs), Zooming Out

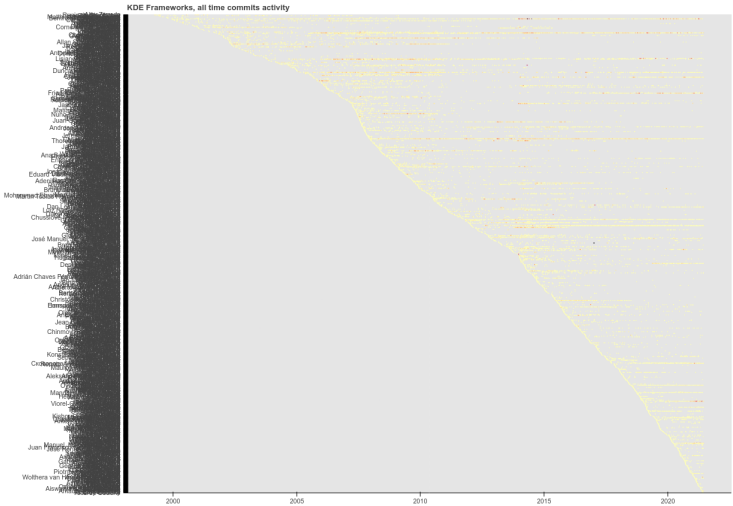


- Krita enters the scene
- Halla Rempt at the center of this subnetwork

Look For Something Hidden

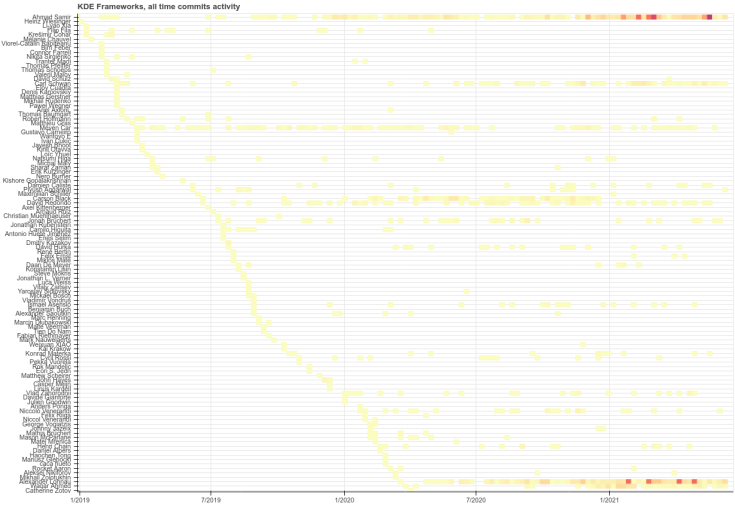
KDE Frameworks

All Time Activity (Commits Only)



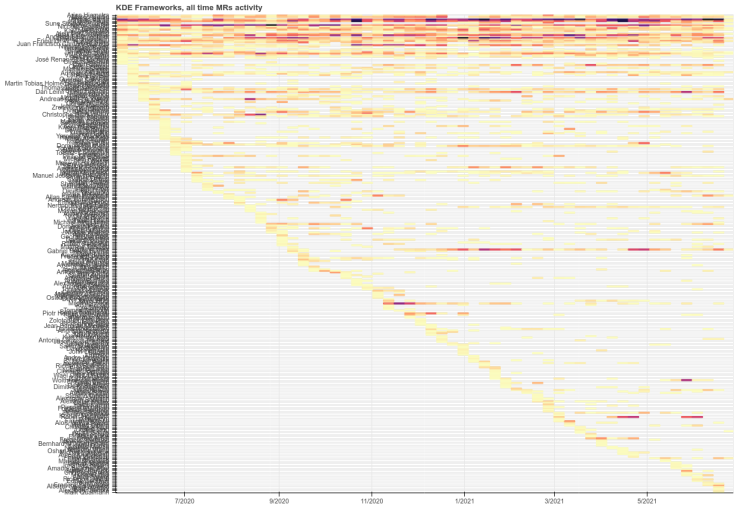
- Very similar profile to the whole KDE
- Two inflexion points though
  - Before 2010, KDE3 / KDE4 transition
  - Around 2014, KDE Frameworks 5

All Time Activity (Commits Only), Recruits Focus



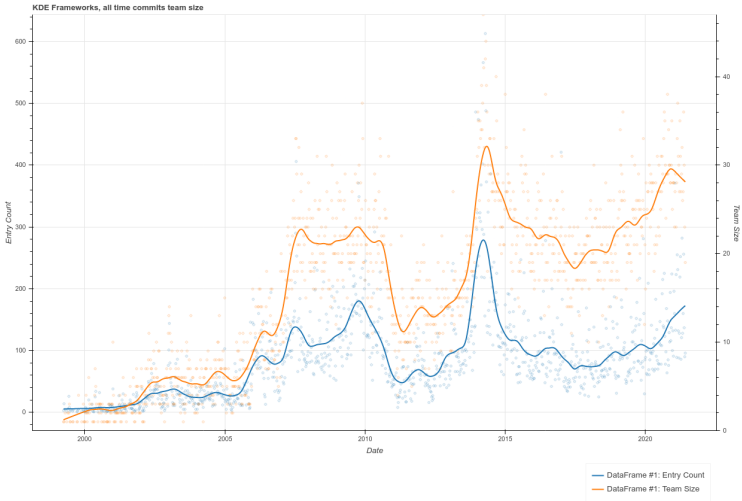
- Ahmad Samir, very end of 2018
- Alexander Lohnau, 2020

All Time Activity (GitLab MRs)



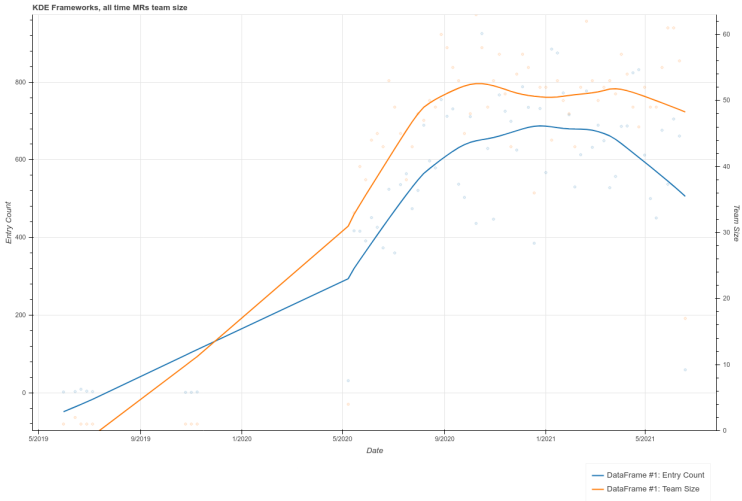
- We see Ahmad Samir and David Faure talking to each other **a lot**

All Time Team Size (Commit Only)



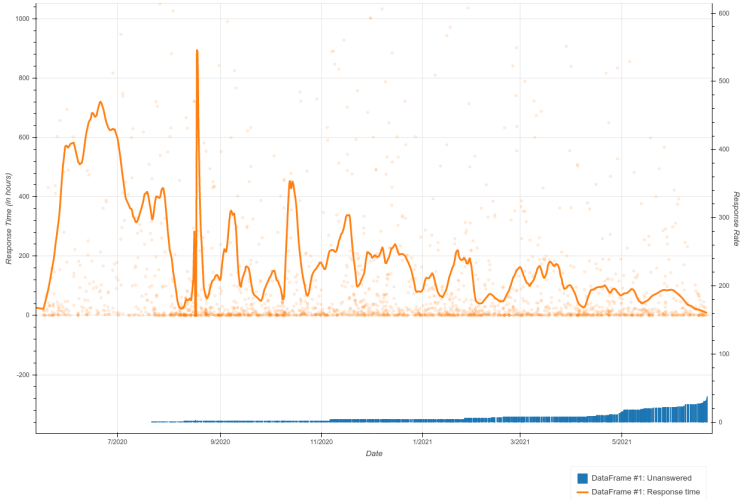
- Ramp up to KDE Frameworks 5 is visible
- Exhaustion phase visible as well
- Picking up quite some pace since a few years
  - Could be development model paying off?
  - Could be preps for KF6?
  - Could be both?

All Time Team Size (GitLab MRs)



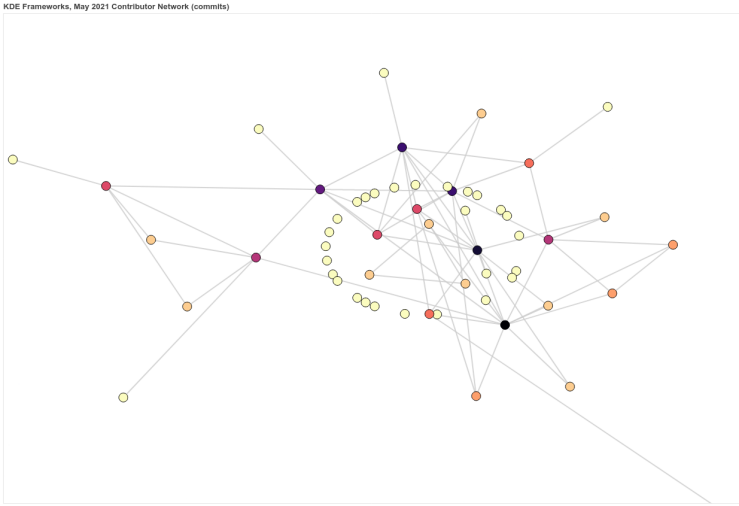
- Huh?
- More data please!

All Time Responsiveness (GitLab MRs)



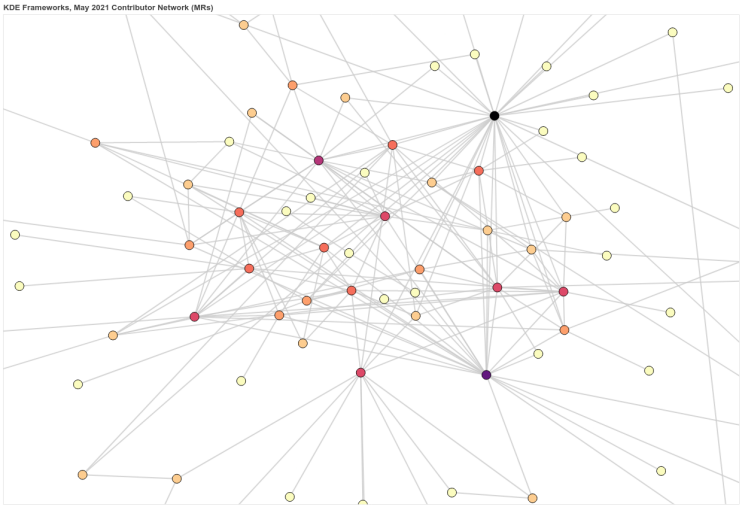
- Similar profile than the whole community
- Slower responsiveness though (3 days to a week on average)

May Network (Commits Only)



- Centrality top 5
  - Ahmad Samir
  - Friedrich Kossebau
  - Alexander Lohnau
  - Laurent Montel
  - Nicolas Fella

May Network (GitLab MRs)



- Centrality top 5
  - Ahmad Samir
  - Nate Graham
  - Nicolas Fella
  - David Faure
  - Alexander Lohnau
- Same effect around Laurent Montel than before
- David Faure is all managerial now

Madness!

Krita

All Time Activity (Commits Only)

- Seems to recruit a bit faster than whole of KDE on average
- Retention seems to be much lower though

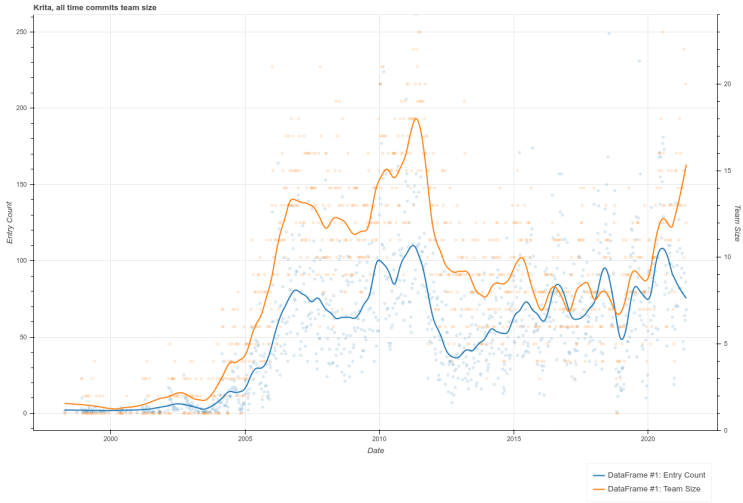
All Time Activity (Commits Only), Recruits Focus

- Eoin O'Neill
- Agata Cacko
- Sharaf Zaman
- Deif Lou

All Time Activity (GitLab MRs)

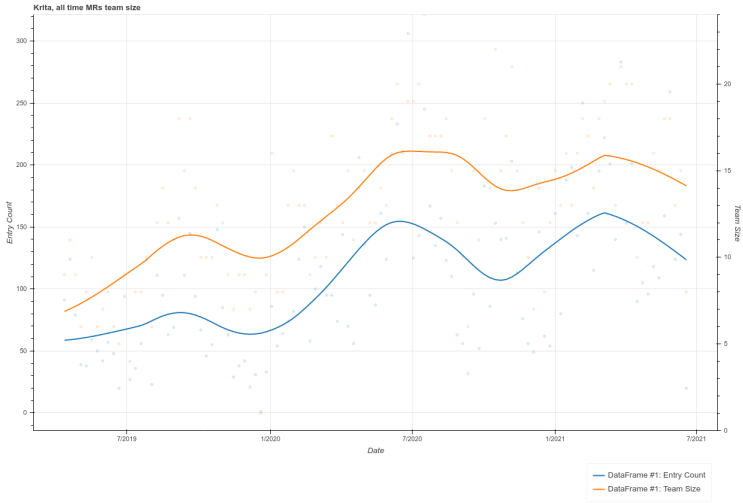
- Dmitry Kazakov
- Halla Rempt

All Time Team Size (Commit Only)



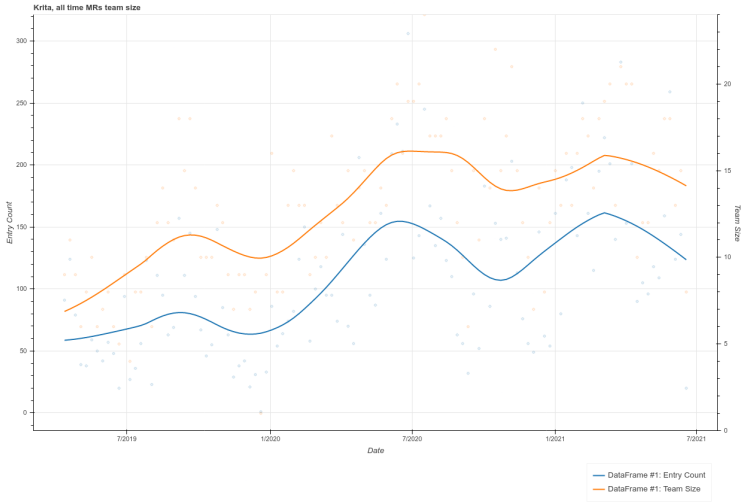
- We see a Nokia peak again
  - Probably coming from when Krita was part of Calligra
- Constantly growing since then
- A bit slower on the team growth

All Time Team Size (GitLab MRs)



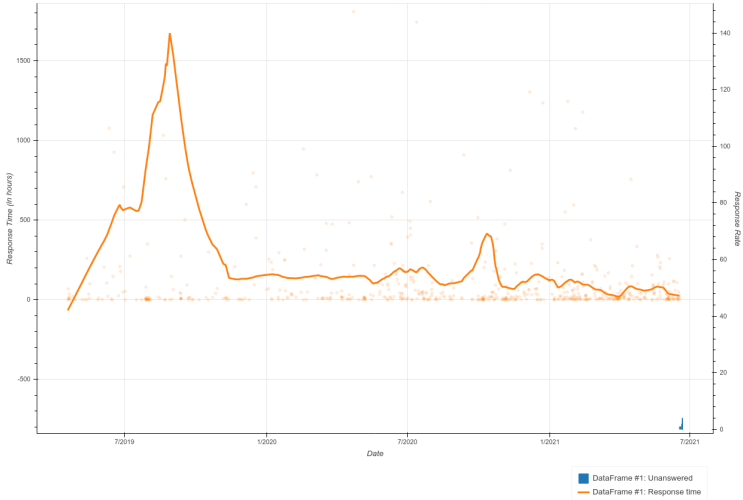
- Did
- We
- Say
- More
- Data?
- Please!

All Time Team Size (GitLab MRs)



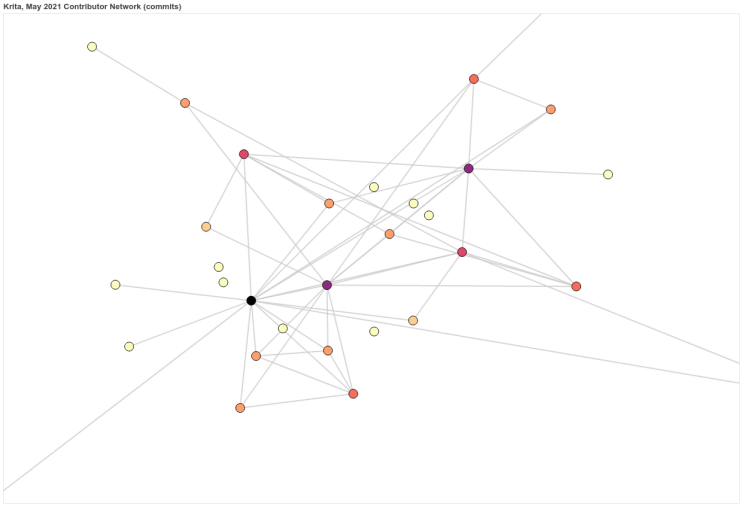
- Did
- We
- Say
- More
- Data?
- Please!

All Time Responsiveness (GitLab MRs)



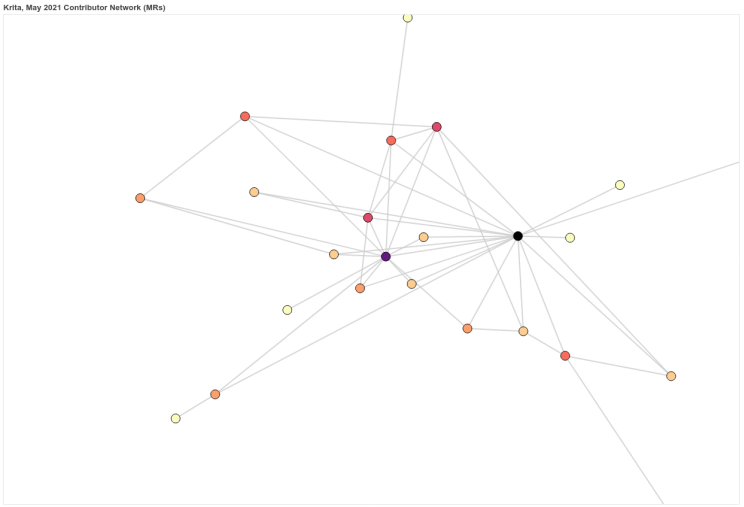
- Similar to the whole KDE community for the responsiveness
- Very tidy on the unanswered stock though
  - Virtually no stock
- Surprising response time peak in 2020 though
  - Unclear why to be honest...
  - Insights welcome!

May Network (Commits Only)



- Centrality top 5
  - Dmitry Kazakov
  - Alvin Wong
  - Halla Rempt
  - Sharaf Zaman
  - Eoin O'Neill

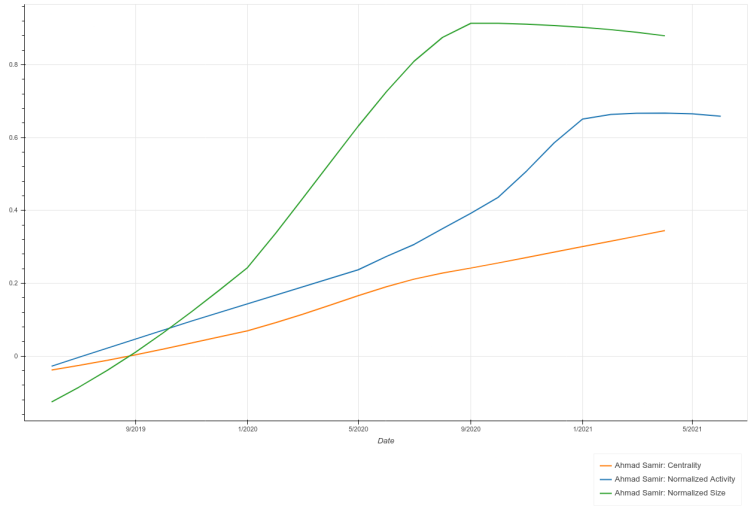
May Network (GitLab MRs)



- Centrality top 5
  - Dmitry Kazakov
  - Halla Rempt
  - Wolthera van Hövell
  - Agata Cacko
  - Alvin Wong

The End is Nigh

A Focus on a Rising Star: Ahmad Samir



- Reminder: only trust values on periods of stable size
- Plot produced on the KDE Frameworks only MRs dataset
- Notice he's just shooting right through it though
- A name to count with if he keeps at it!

Conclusion

The Metrics

- Interesting to have at last the conversations during code reviews
  - Allows to capture a bit how senior developers evolve
- Commits are probably not the best way to produce contributor networks
  - To be used only as a fallback if code reviews data is unavailable
- Important to combine visualizations to get a proper view at a community
  - A single metric gives too much bias to analyses

The Community

- Complex history with a couple of decompression phases visible
- Healthy overall
  - Recruiting going well
  - Retention seems to pick up as well
- Subprojects have their own trends
  - Fairly close to the whole
    - At least where we looked
- Clearly GitLab is big for KDE already
  - Looking forward to the accumulated data in a few years

The Future?

- Still work needed on the responsiveness plot
- Find ways to reliably match GitLab accounts and commit authors
  - Opens the door to combined commits/MRs views
- Have a way to customize weights on activities or network edges

Goodbye!

Thanks You!

Questions?

christelle.zouein@enioka.com

ervin@kde.org

kevin.ottens@enioka.com