Gender-diversity analysis of technical contributions

LinuxCon, Berlin 2016

Daniel Izquierdo Cortázar
@dizquierdo
dizquierdo at bitergia dot com
https://speakerdeck.com/bitergia

Outline

Introduction

First Steps

Some numbers and method

Conclusions

Introduction

A bit about me

Why this analysis

What we have so far

/me

CDO in Bitergia, the software development analytics company

Lately involved in understanding the gender diversity in some OSS communities

Involved in OPNFV dashboard (opnfv.biterg.io)

Disclaimer: *not involved in any working group, own analysis and interest, I may have missed some stuff...*

Why this study

Diversity matters

I attended some (Women of OpenStack) talks in the OpenStack Summit (Tokyo and Austin)

There are not numbers about technical contributions (AFAIK)

Produced some numbers that gained some attention, so this is for sure of interest for the Linux ecosystem

In the end this is all about **transparency** and improvement

What we have so far

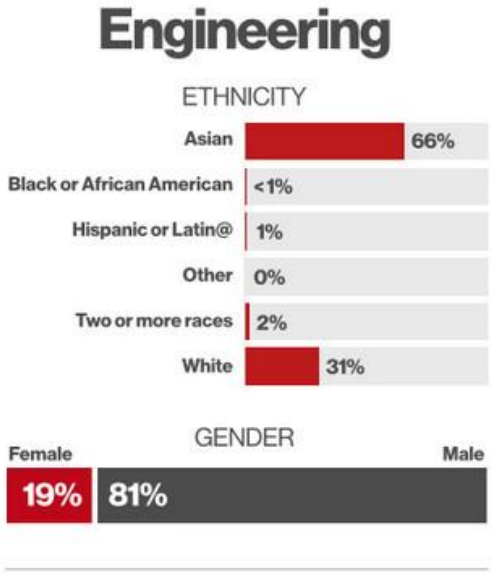
FOSS Survey in 2013:

- http://floss2013.libresoft.es/results.en.html
- 11% of women answered the survey

The Industry Gender Gap by the World Economic Forum.

- 5% for CEOs, 21% for Mid-level roles, 32% of Junior roles

Some companies



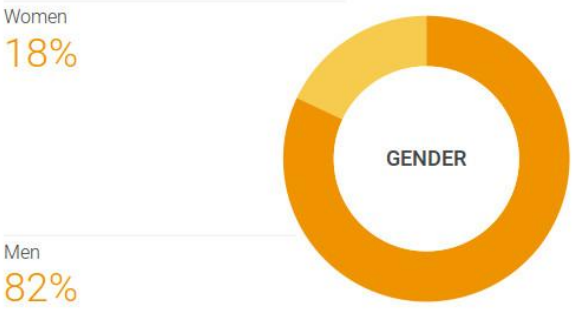
Pinterest Engineering focused employees.

https://blog.pinterest.com/en/our-plan-more-diverse-pinterest

Bitergia

Some companies

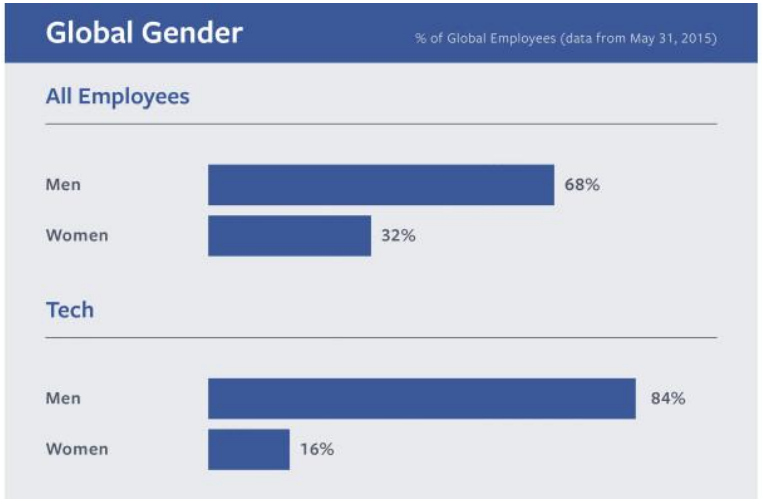
Google Tech focused employees.



http://www.google.com/diversity/

Bitergia

Some companies



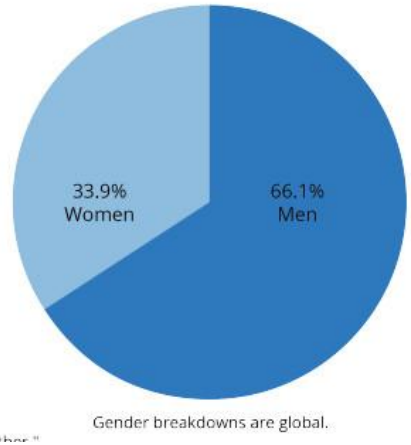
Facebook Tech focused employees.

http://newsroom.fb.com/news/2015/06/driving-diversity-at-facebook/

Bitergia

Some companies

Dropbox all employees.

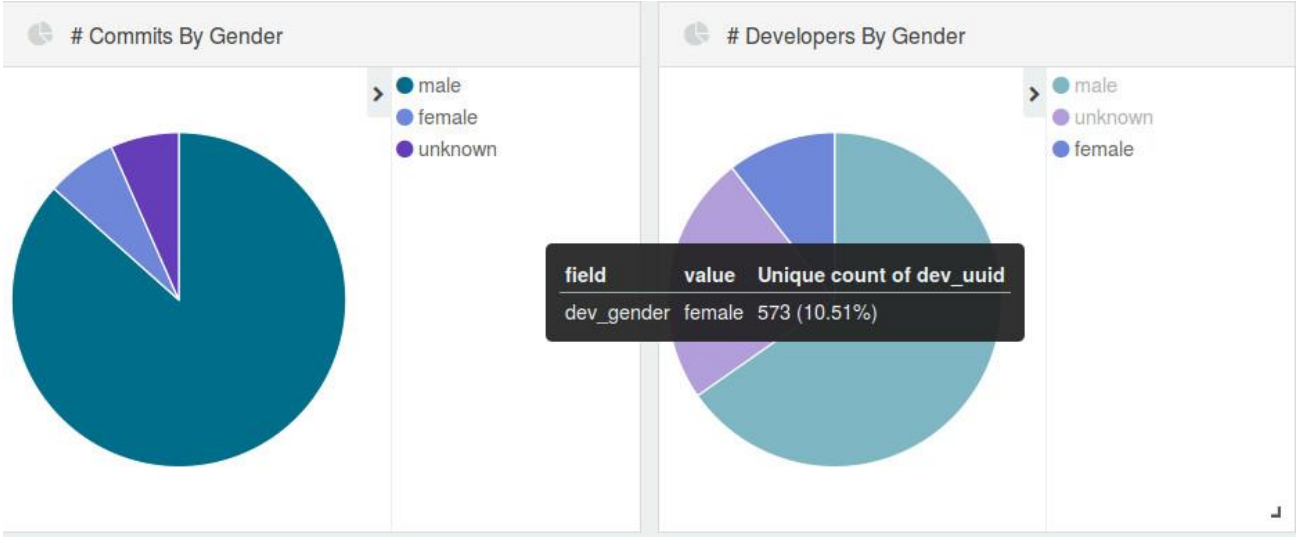


https://blogs.dropbox.com/dropbox/2014/11/strengthening-dropbox-through-diversity/

Bitergia

OpenStack numbers

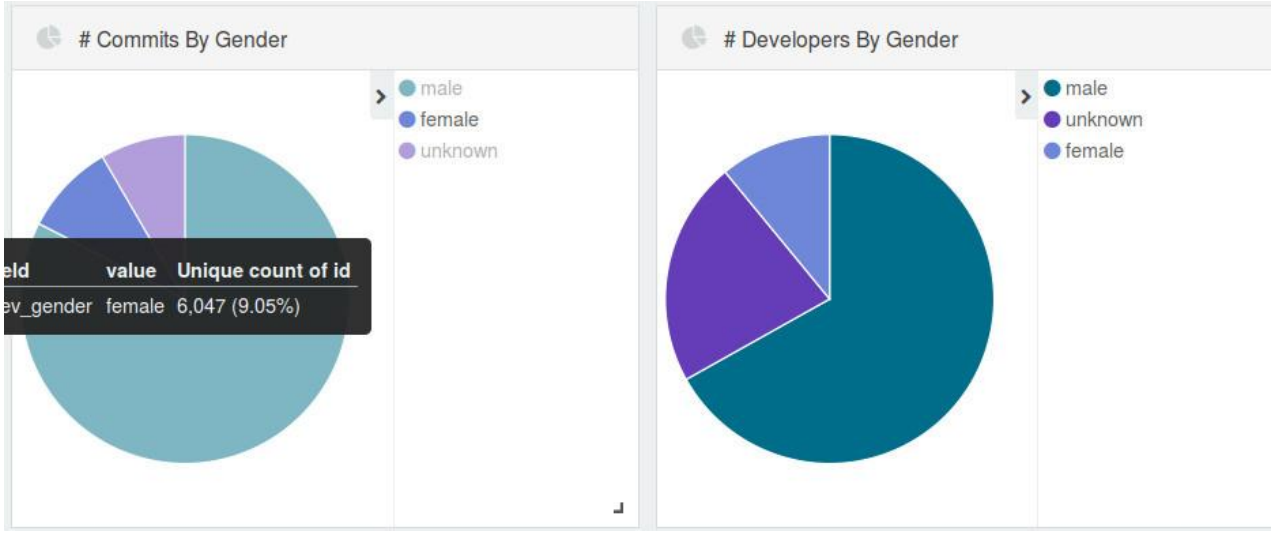
Women activity (**all of the history**):
~ 10,5% of the population (~ 570 developers)
~ 6,8% of the activity (>=16k commits)



Bitergia

OpenStack numbers

Women activity (**last year**):
~ 11% of the population (~ 340 active developers)
~ 9% of the activity (>=6k commits)



Bitergia

Summary

Conclusions not representative, but:

- Women represents around 30%/40% of the workforce in tech companies.
- And between 10% and 20% if focused on tech teams.
- OpenStack shows a 11% of the population
- What about the Kernel?

First Steps

Some Definitions

Technical contributions: commit, flag in the mailing list (acked-by, reviewed-by), email related to the code review

Other potential metrics: diversity by company, fairness in the code review among organizations and genders, transparency in the process

Available but sensitive info: affiliation, countries, time to review

First Steps

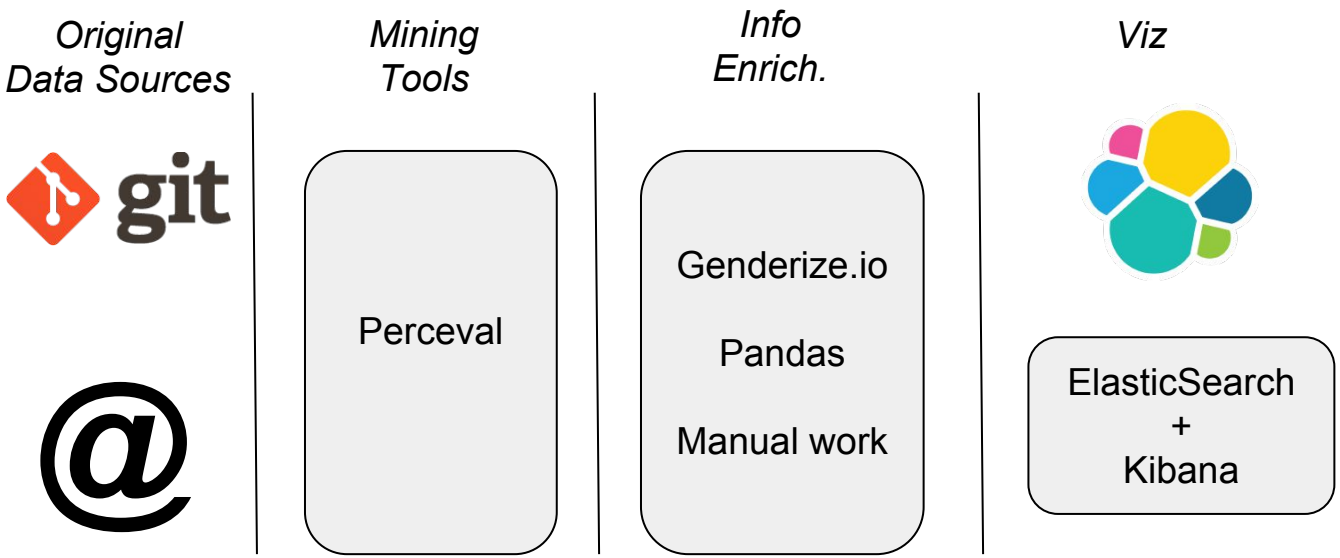
Names databases

Genderize.io

Manual analysis

Focus on main developers

Architecture



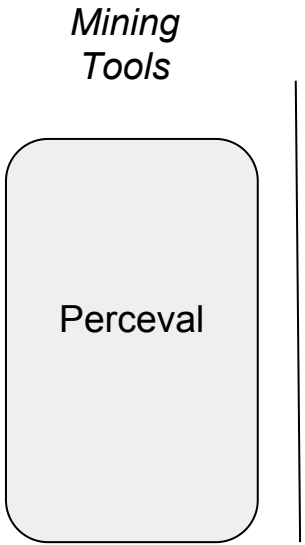
Bitergia

Architecture

Original Data Sources

- Git and mailing lists
- ~ 600k commits (starting in 2006)
- ~ 3.8M emails
- ~ 1.4M emails with keyword PATCH
- ~ 2.5M tags

Architecture



- Produces JSON documents from the usual data sources in OSS
- Part of the GrimoireLab toolchain
- grimoirelab.github.io

Bitergia

Architecture

Info
Enrich.

Genderize.io

Pandas

Manual work

- Genderize.io: name database
- Pandas: data analysis lib.
- Ceres library (dicortazar/ceres @ github)
- Manual work:

Bitergia

Architecture

Viz

ElasticSearch
+
Kibana

- ElasticSearch: Schemaless db
- Kibana: works great with ES
- This tandem helps a lot to verify info
- Drill down capabilities
- Extra info available (but not displayed)

Bitergia

Validation: manual work

Check main contributors by hand

Asian names hard to check (u_u) (help needed!)

Lack of mailing lists (gmane service ended)

Outreachy names successfully added to the analysis (only 3 of them were wrongly assigned by the API)

Some numbers

Git Contributions

Mailing List Activity

Demographics

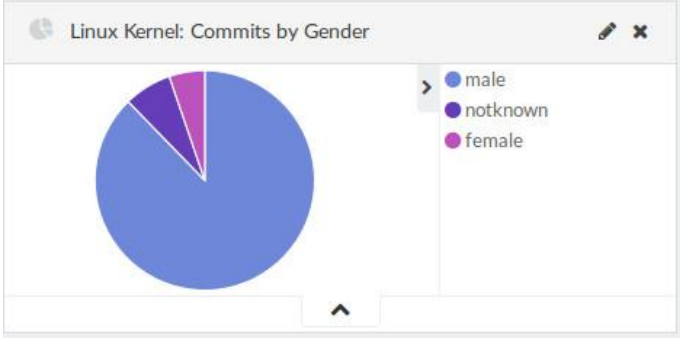
Git Overview

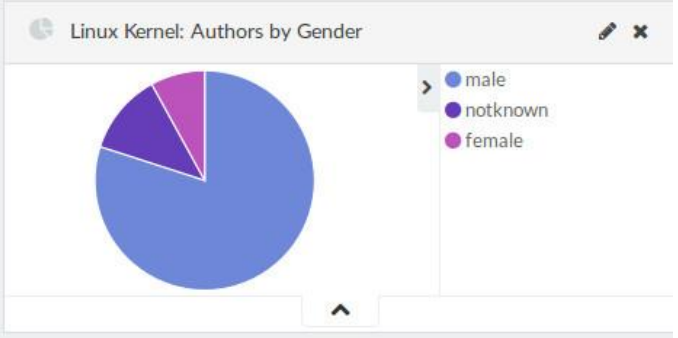
- Aggregated historical data
- Linus Torvalds GitHub Git repository

Bitergia

Git Activity and Population

Women activity (**since 2005)**:
~ 5.2% (> 31K commits)
~ 8% of the population (~ 1,15K developers)

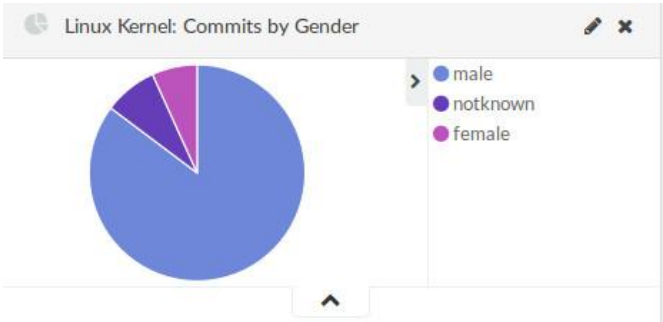


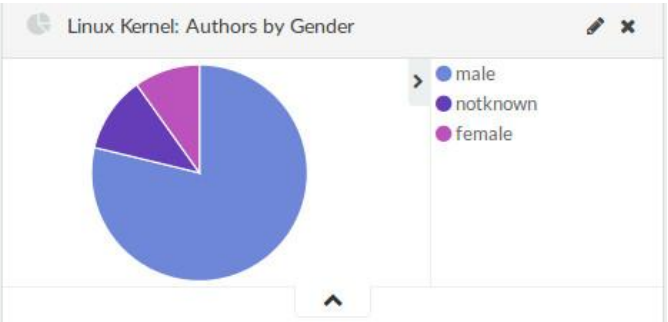


Bitergia

Git Activity and Population

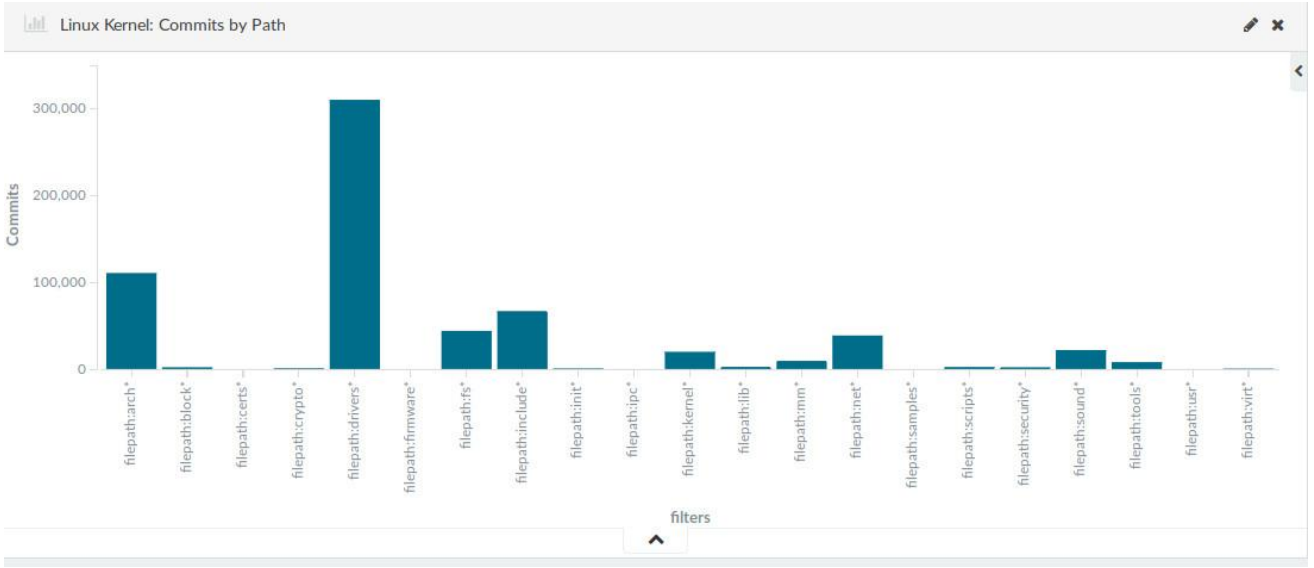
Women activity (**last year**):
~ 6.8% of the activity (~ 4k commits)
~ 9.9% of the population (~ 330 active developers)





Bitergia

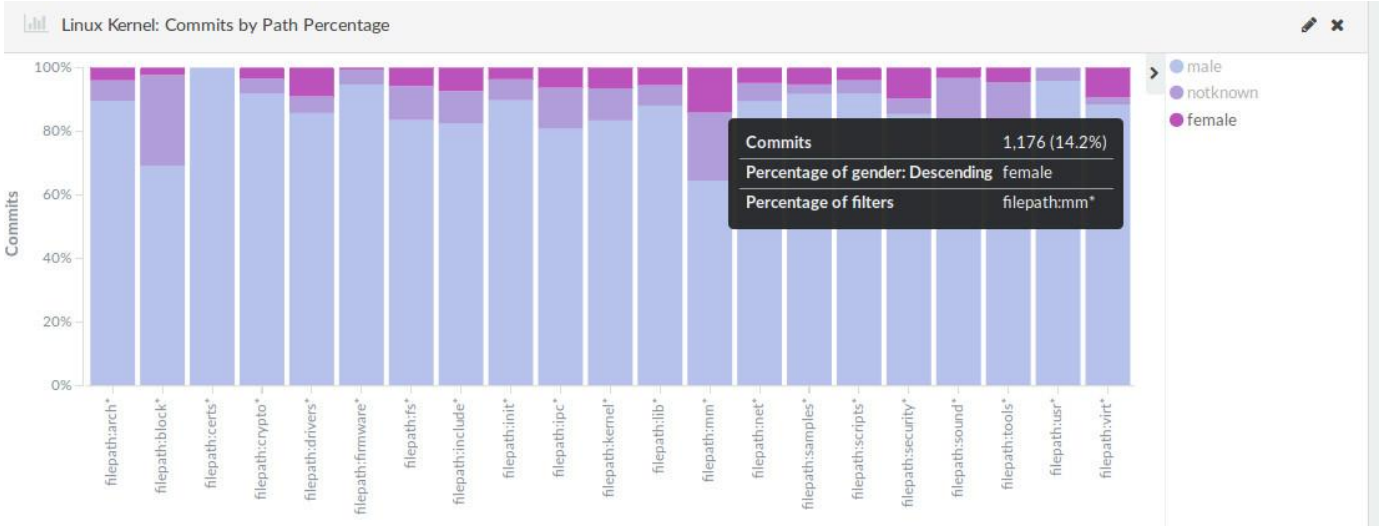
Git Main Modules



Arch and drivers are the most active directories with contributions

Bitergia

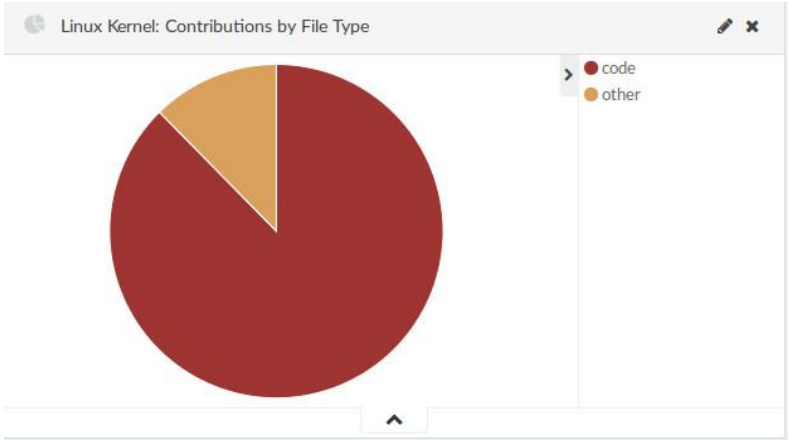
Git Main Modules



Bitergia

Drivers (~10% of activity) and mm (~15% of activity) directories the most diverse

Git Type of Contribution

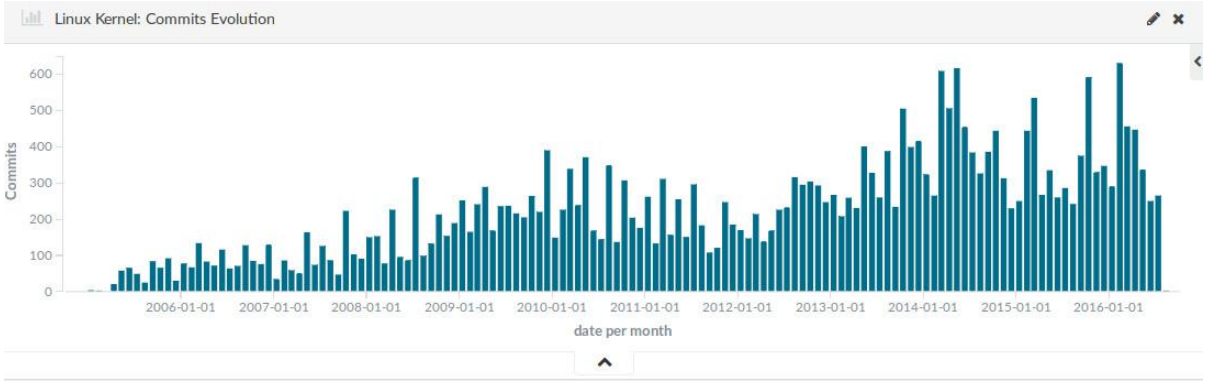


- Code: .c, .h, .cpp, etc
- Other: Makefile, .txt, etc
- 87% of contributions are code.
- Women are over the mean with >= 90%

Bitergia

Git Activity Women Evolution

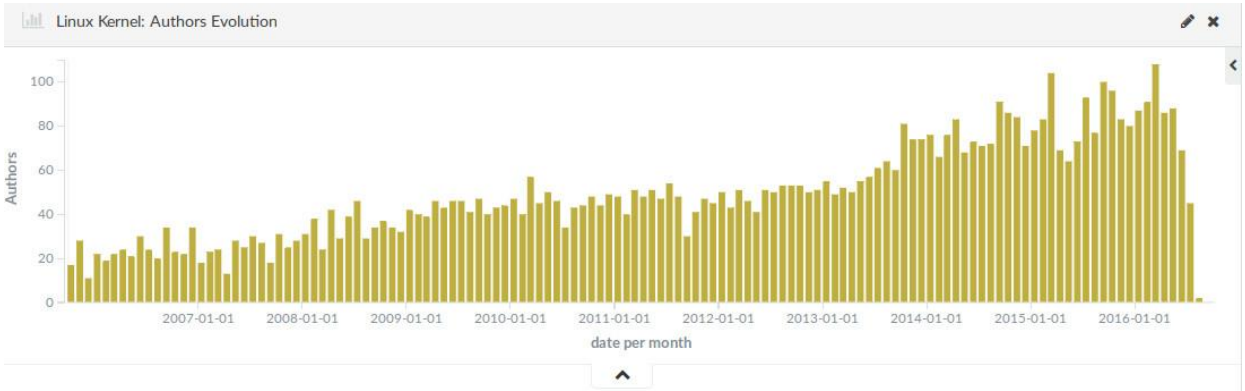
- Similar trend than the overall evolution
- Peaks during mid 2014 and mid 2016 (any clue?)



Bitergia

Git Authors Women Evolution

- Small jump in 2014
- More contributors since then (any clues?)



Bitergia

Mailing Lists Overview

Linux Kernel mailing list

Flags = Tags = [Reviewed-by|Acked-by|Signed-off-by|...]

Gender analyzed for the email sender and in the flags/tags

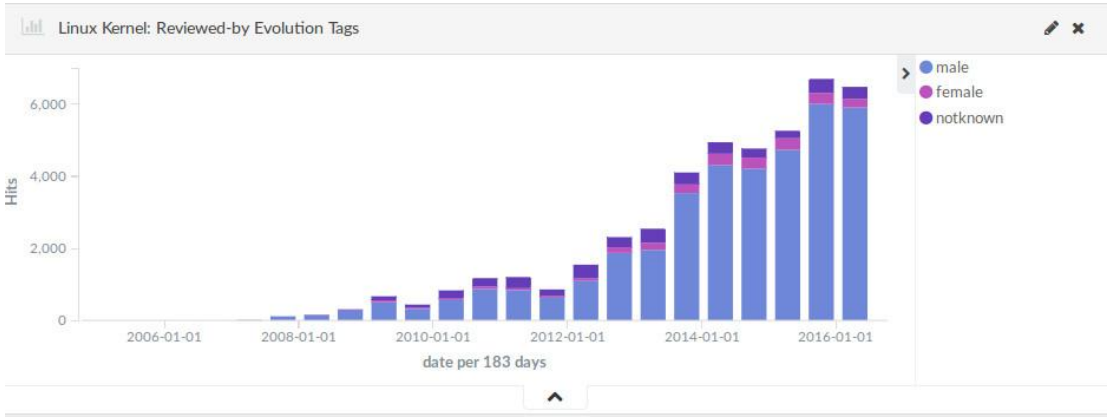
Bitergia

Code Reviews (Reviewed-by)

2014 Activity Jump: more complex processes? Longer reviews?

Jump also seen when splitting by men or women

Reviewed-by by women between 4% and 6%



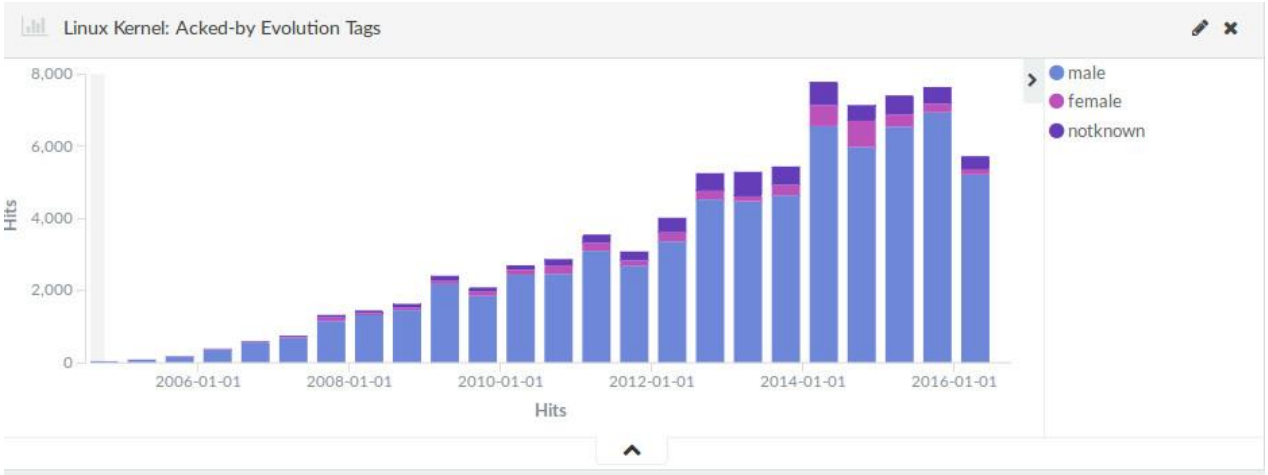
Bitergia

‘Merging’ Code Reviews (Acked)

2014 not-that-big Jump

Jump also seen when splitting by men or women

Acked-by by women between 3% and 10%



Bitergia

Demographics

Attraction of female developers to the community
Peak on 2014/2015 with up to 110 developers

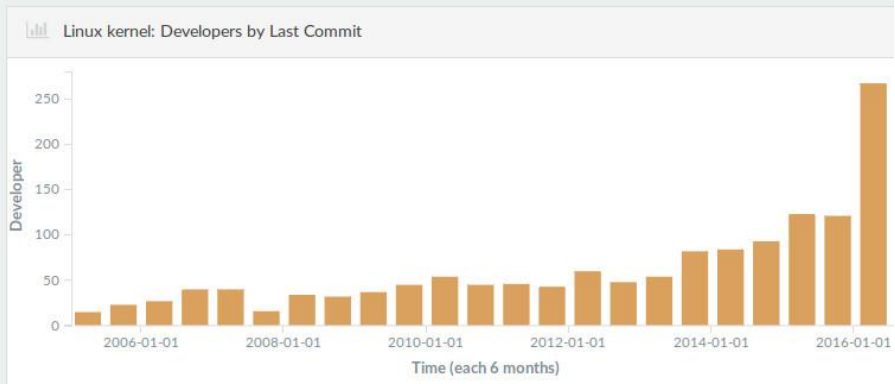
[chart measures the first contribution by each developer and groups by six months]

Bitergia

Demographics

Female developers leaving the community

[active developer = at least a commit during the last year]
[chart measures the last contribution by each developer and groups by six months]



Bitergia

Demographics: extra bonus

When were born the developers contributing during the last quarter?

And who are they? Working for? Working at?

Bitergia

Demographics: extra bonus

And the other way around:

How good are we retaining developers that entered in 2013-S1?
(And who are they? Working for? Working at?)

[*64 attracted in 2013 S1. 35 left in that quarter. 12 are still contributing. Another 17 left in other periods*]

Bitergia

Analysis

Comparison with the OpenStack Community

Comparison

Let’s have in mind:

- Different process to code review
- Different mission
- Different programming language
- Different governance
- 1 project vs N
- <Add here your favourite difference!>

Comparison

But:

- Continuous increase of women attracted in both cases (11% vs 10% in the Kernel)
- Jump in contributors in the case of the Kernel
- Jump in code review process in the case of OpenStack

Conclusions

Answer to First Questions

Data to Make Decisions

Open Paths

Some Answers

Continuous increase of activity and population (up to 10%)

Remarkable increase in Git population after 2014

Tooling is useful to have numbers, compare and make decisions or check policies

Others: the code review seems to be increasing its activity (reason for 2014 jumps in activity? -> this may lead to more noise)

Conclusions

Room for improvement of the dataset

This provides some initial numbers about the current status

Hopefully useful for the Foundation and the Kernel project itself

Potential Actions

How this may help some challenges when attracting women:

- Close to 1110 female developers (more than 400 with a 100% of probability)
- Talk to them, send an email, let them participate, have meetings, ask for mentorships
- Detection of new women entering the community, say hello!

Further Work

Sensitive info: dashboard still private

Extra analysis: time to merge **fairness**, **companies** women %, **Outreachy** follow ups, **quarterly** reports, updated data, specific policies **ROI** and others.

This [hopefully] helps to have a better picture

Other minorities analysis could be done

Bitergia

How can you help?

Is there a formal working group focused on women in the Linux Foundation/Kernel?

Have you defined policies in this area?

Are there good practices to create safe and productive environments?

Looking for sponsors!

Gender-diversity analysis of technical contributions

LinuxCon, Berlin 2016

Daniel Izquierdo Cortázar
@dizquierdo
dizquierdo at bitergia dot com
https://speakerdeck.com/bitergia