The Comment Density of Open Source Software Code

Oliver Arafat
Siemens AG, Corporate Technology
oarafat@gmail.com

Dirk Riehle
SAP Research, SAP Labs LLC
dirk@riehle.org, http://dirkriehle.com
http://twitter.com/dirkriehle

Oliver Arafat, Dirk Riehle. “The Comment Density of Open Source Software Code.” In ***Companion to Proceedings of the 31st International Conference on Software Engineering*** (ICSE 2009). IEEE Press, 2009. Page 195-198.

[Link: http://dirkriehle.com/2009/02/04/the-comment-density-of-open-source-software-code/]

Overview

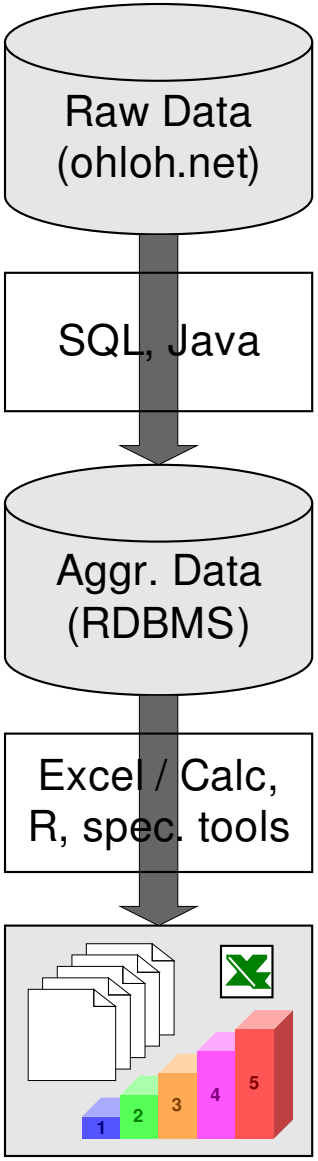
Summary (blog post): The Sweet Spot of Code Commenting in Open Source

[Link: http://dirkriehle.com/2009/02/04/the-sweet-spot-of-code-commenting-in-open-source/]

Abstract: The development processes of open source software are different from traditional closed source development processes. Still, open source software is frequently of high quality. Thus, we are investigating how open source software creates high quality and whether it can maintain this quality for ever larger project sizes. In this paper, we look at one particular quality indicator, the density of comments in open source software code. In a large-scale study of more than 5,000 projects, we find that active open source projects document their source code, and we find that the comment density is independent of team and project size, but not of project age. In future work, we intend to correlate comment density with project success or failure.

Reference: Oliver Arafat, Dirk Riehle. "The Comment Density of Open Source Software Code." In ***Companion to Proceedings of the 31st International Conference on Software Engineering*** (ICSE 2009). IEEE Press, 2009. Page 195-198. [Link: http://dirkriehle.com/2009/02/04/the-comment-density-of-open-source-software-code/]

Analysis Process and Tool Chain



Raw data source

- Local database (ohloh.net snapshot, crawled sources)
- Web services access (ohloh.net, sourceforge.net, others)

Pre-processing

- Database querying using SQL and scripts
- Java library for computationally heavyweight filters, aggregation

Aggregated data source

- Output of pre-processing stage for specific analytical tasks
- Aggregated data significantly improves analysis speed

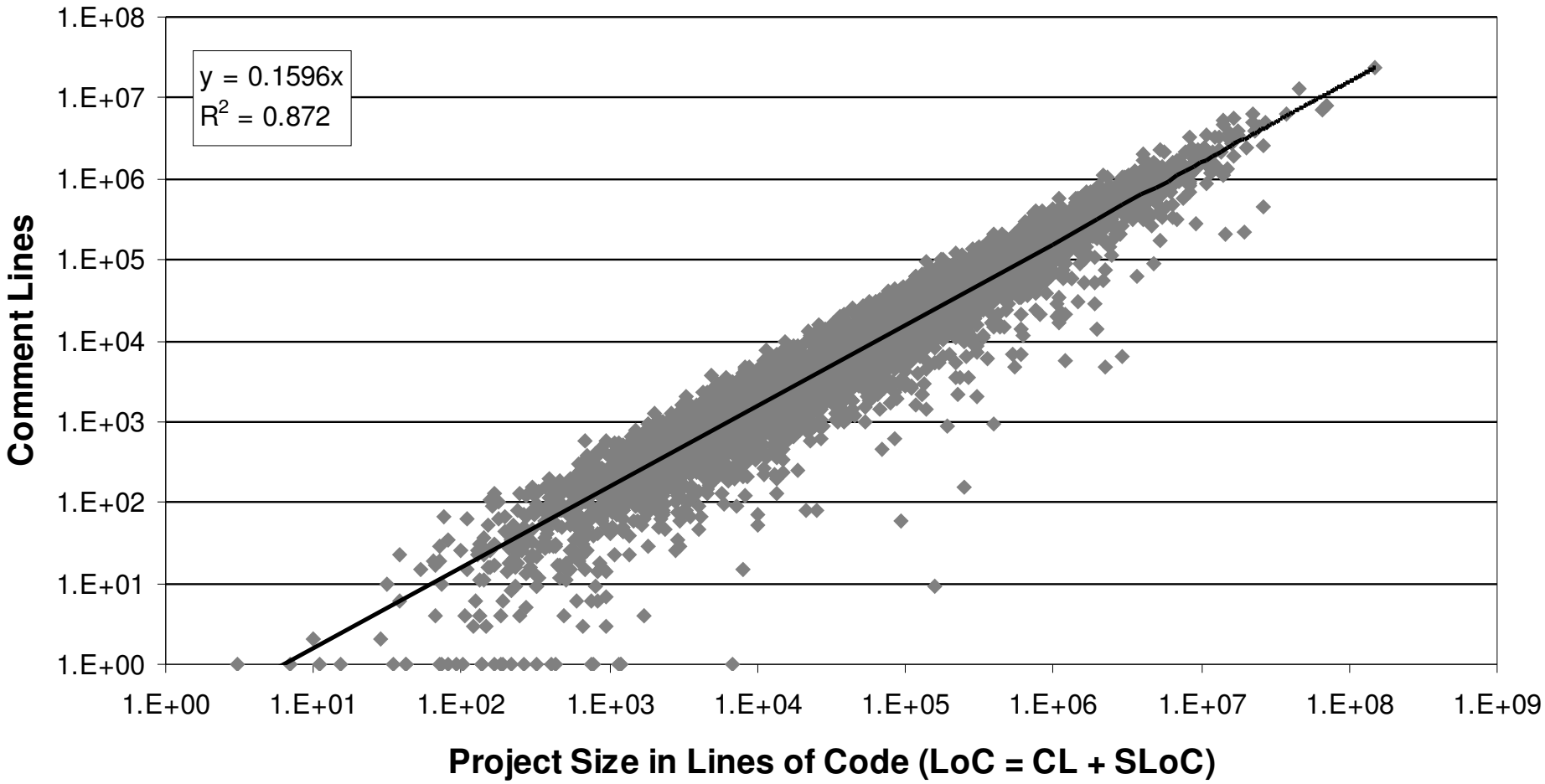
Analytical processing

- Mines aggregated (and raw) data for insights, hypothesis testing
- At present basic processing (Excel), machine learning next

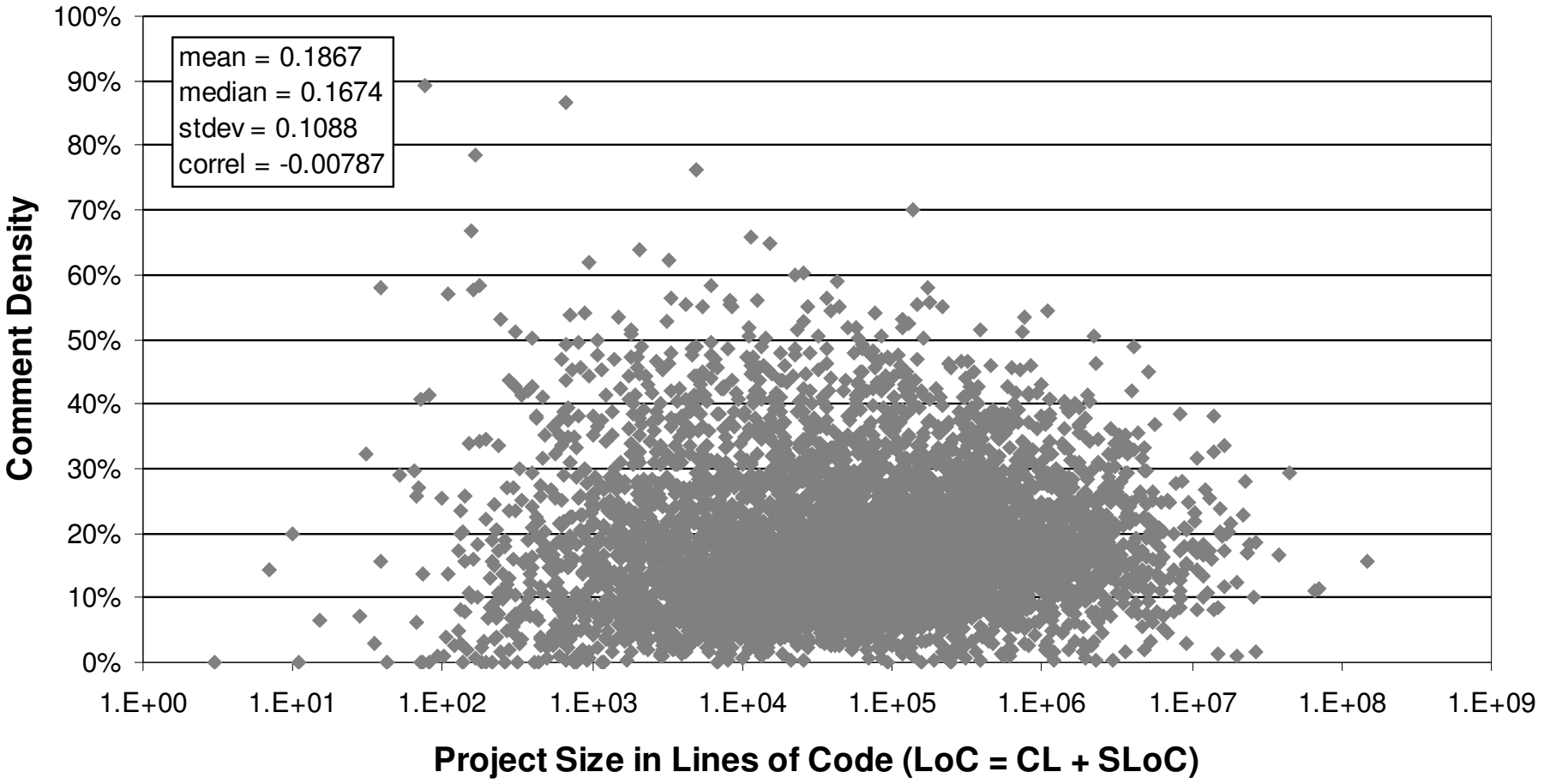
Analysis output

- Results of analytical processing: averages, distributions, correlations
- Presented as models, tables, graphs, charts, etc.

Commenting Practice in Open Source



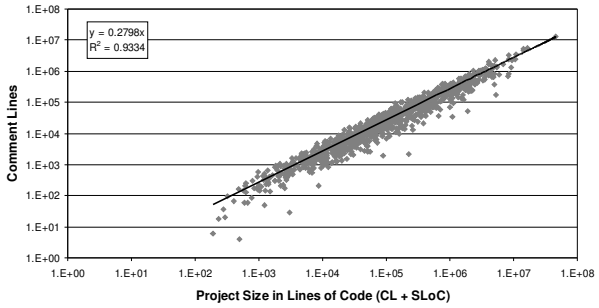
Comment Density

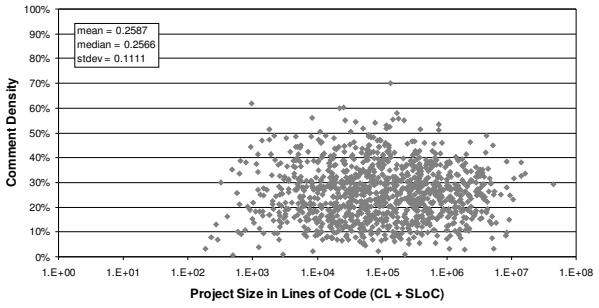


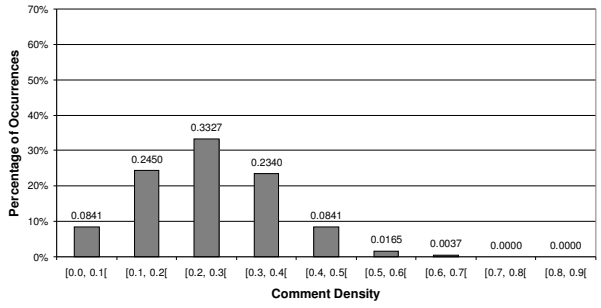
- **Comment density**
  - Percentage of lines of code that are comments
  - Measure of how well documented code is
  - Indicator of likelihood of survival
- **Definitions**
  - CL = comment lines
  - SLoC = source lines of code
  - CD = CL / (CL + SLoC)

Variation by Programming Language

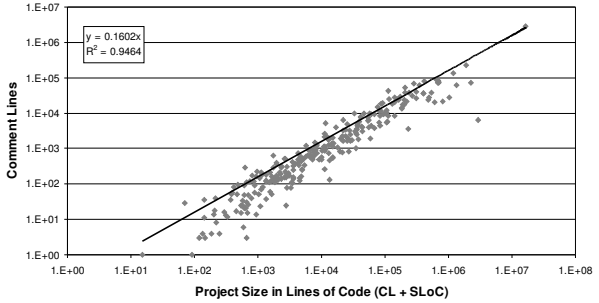
Java

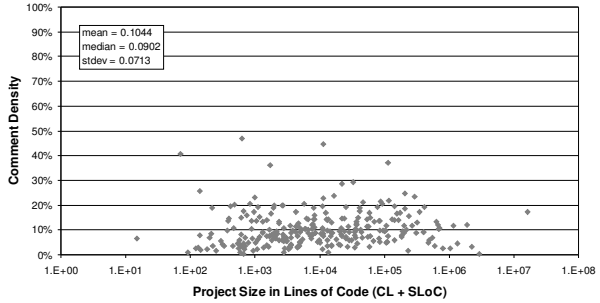


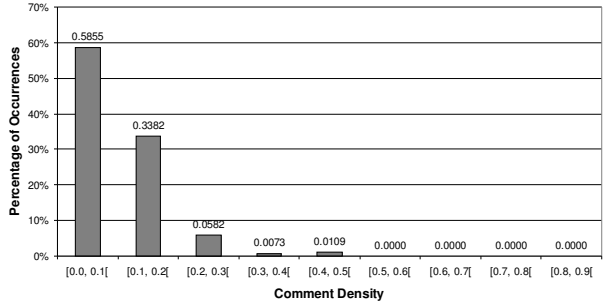




PERL



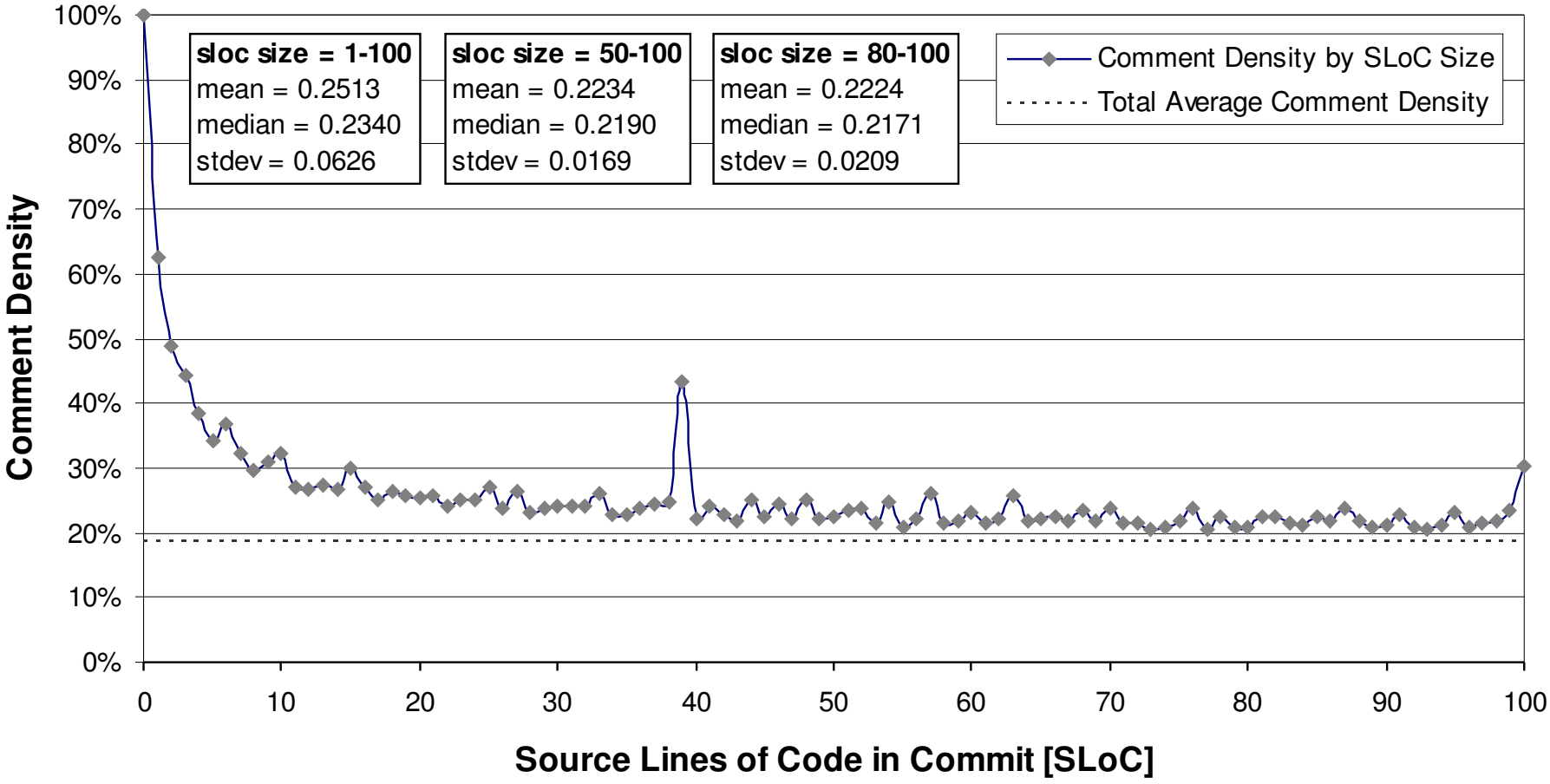




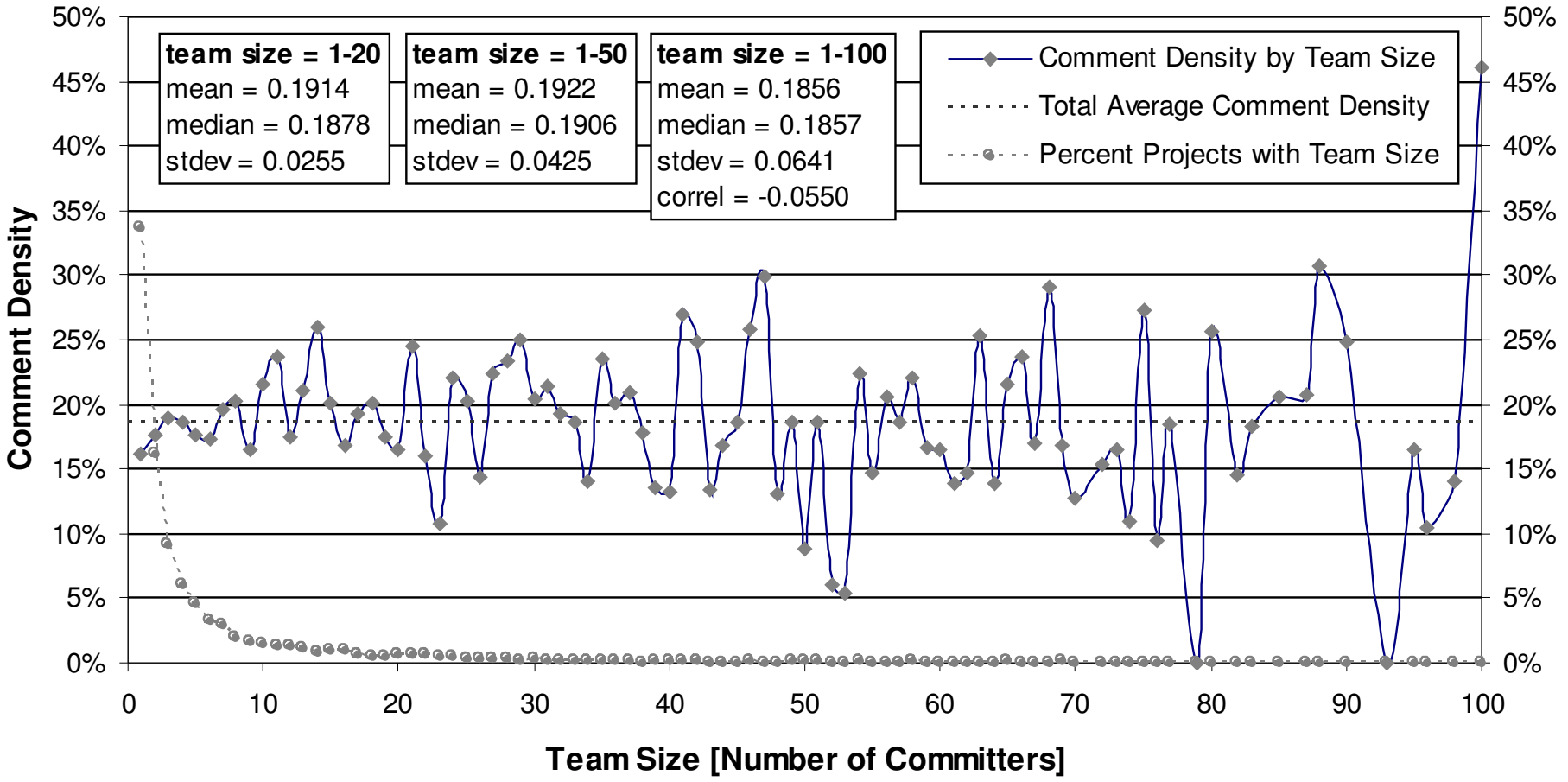
More here: How Open Source Comments (by Programming Language)

[Link: http://dirkriehle.com/2008/11/10/how-open-source-comments-by-programming-language/]

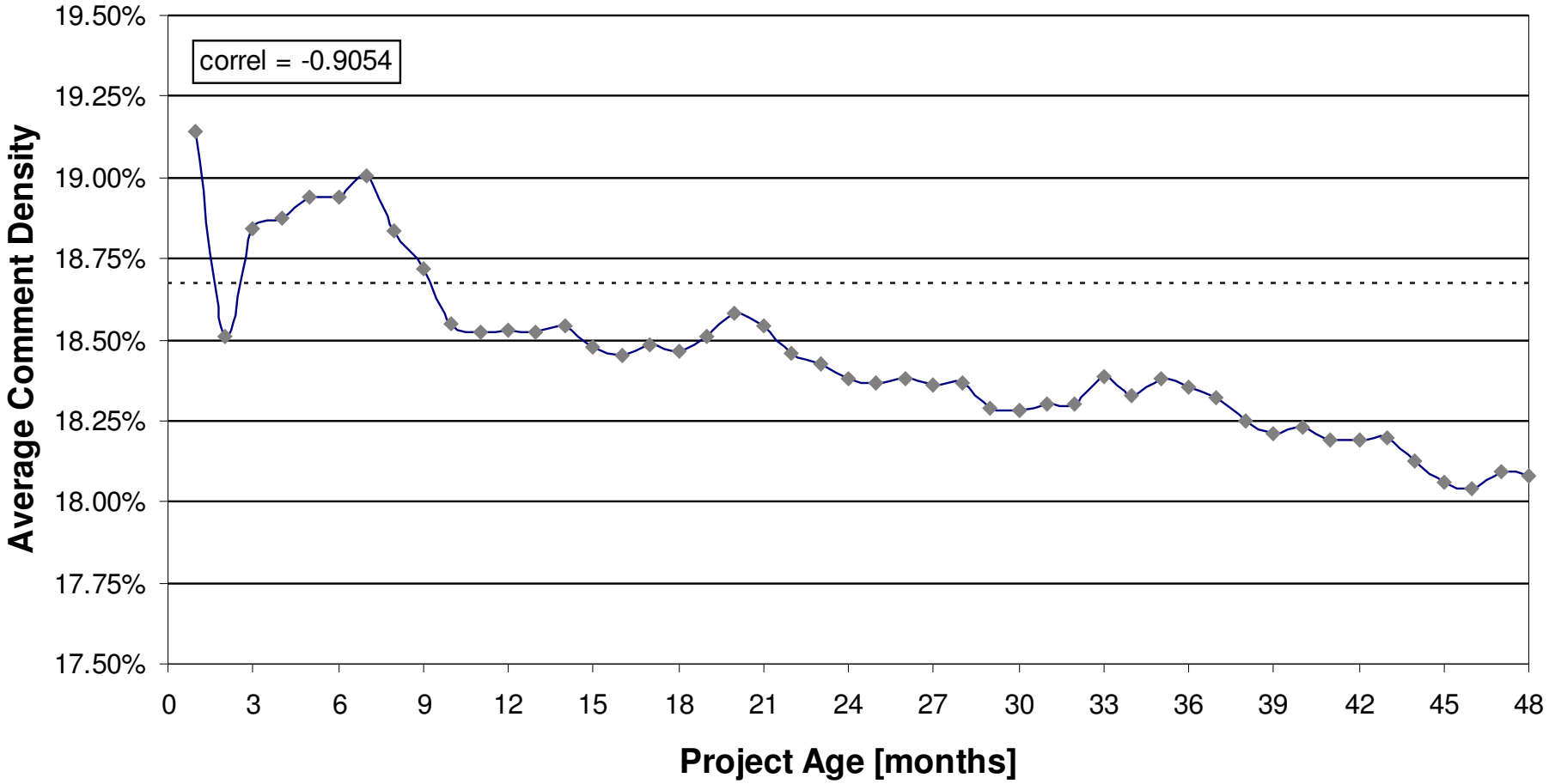
Comment Density by Commit Size



Comment Density by Team Size



Comment Density by Project Age



Commenting Practice Summary

- Successful open source projects follow a consistent commenting practice
  - 1 out of 20 commits serves commenting purposes only
  - 1 out of 5 non-empty lines is a comment line
- Average comment density is independent of project size
- Average comment density is independent of team size
- Average comment density slowly falls with project age

References

- Dirk Riehle, John Ellenberger, Tamir Menahem, Boris Mikhailovski, Yuri Natchetoi, Barak Naveh, Thomas Odenwald. "Bringing Open Source Best Practices into Corporations Using a Software Forge." ***IEEE Software***, 2009. See: http://dirkriehle.com/2009/02/11/open-collaboration-within-corporations-using-software-forges/
- Dirk Riehle. "The Economic Motivation of Open Source: Stakeholder Perspectives." ***IEEE Computer***, vol. 40, no. 4 (April 2007). Page 25-32. See: http://dirkriehle.com/computer-science/research/2007/computer-2007.html
- Oliver Arafat, Dirk Riehle. "The Commit Size Distribution of Open Source Software." In ***Proceedings of the 42nd Hawaiian International Conference on System Sciences*** (HICSS 42). IEEE Press, 2009. See: http://dirkriehle.com/2008/09/23/the-commit-size-distribution-of-open-source-software/
- Amit Deshpande, Dirk Riehle. "The Total Growth of Open Source." In ***Proceedings of the Fourth Conference on Open Source Systems*** (OSS 2008). Springer Verlag, 2008. Page 197-209. See: http://dirkriehle.com/2008/03/14/the-total-growth-of-open-source/
- Amit Deshpande, Dirk Riehle. "Continuous Integration in Open Source Software Development." In ***Proceedings of the Fourth Conference on Open Source Systems*** (OSS 2008). Springer Verlag, 2008. Page 273-280. See: http://dirkriehle.com/2008/03/08/continuous-integration-in-open-source-software-development/
- Philipp Hofmann, Dirk Riehle. "Estimating Commit Sizes Efficiently." In ***Proceedings of the 5th International Conference on Open Source Systems*** (OSS 2009). Springer Verlag, 2009. Page 105-115. See: http://dirkriehle.com/2009/02/11/estimating-commit-sizes-efficiently/
- Oliver Arafat, Dirk Riehle. "The Comment Density of Open Source Software Code." In ***Companion to Proceedings of the 31st International Conference on Software Engineering*** (ICSE 2009). IEEE Press, 2009. Page 195-198. See: http://dirkriehle.com/2009/02/04/the-comment-density-of-open-source-software-code/

Thank you! Questions?

For any feedback or questions, please email the authors:

Oliver Arafat
Siemens AG, Corporate Technology
oarafat@gmail.com

Dirk Riehle
SAP Research, SAP Labs LLC
dirk@riehle.org, http://dirkriehle.com
http://twitter.com/dirkriehle