

基于开发日志的ASOP开发实践的度量和预测

姓名：韩慧阳

学号：2012011276

导师：高性能所-陈康老师、软件所-陈渝老师

提纲

目标重述

论文研究进展

个人进展

后期安排

目标重述

- 根据openthos (in github)、 android-x86(in sourceforge)、 AOSP的开发人员的gitlog日志文件/email等信息，分析人员与项目具体feature的关系，人员的活跃度等。
- 调整：聚焦在linux kernel上，通过git log 信息对人员开发情况进行分析，通过benchmark和性能数据对linux kernel的性能特征和变化情况进行分析

论文研究进展

- 研究背景的调研 开源软件特性去留的关键影响因素:
 - 非性能缺陷 (传统系统缺陷)
 - 可以被严格检测
 - 容易捕捉
 - 检测方法成熟
 - 性能缺陷 (Performance Bugs)
 - 不能被传统方式检测
 - 多数不影响系统的正确性
 - 强烈影响用户体验和软件表现

论文研究进展

性能缺陷的存在会影响到一个新的特性能否成功发布以及旧特性是否能继续留下 性能缺陷的检测、预测对于软件最终发行意义重大

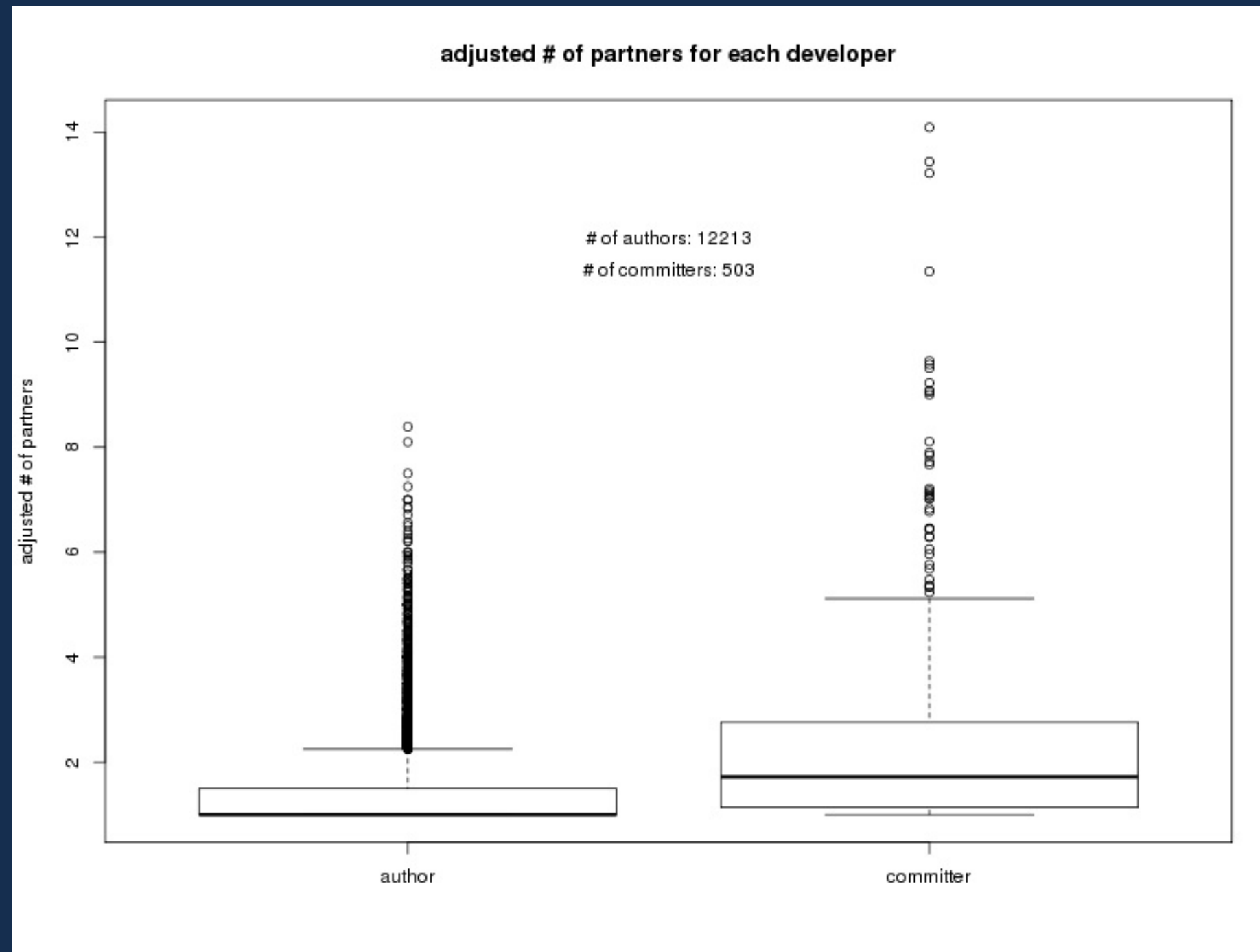
- 主要的性能缺陷检测方法：
 - lkp-compare
 - Toddler
 - cache

个人进展

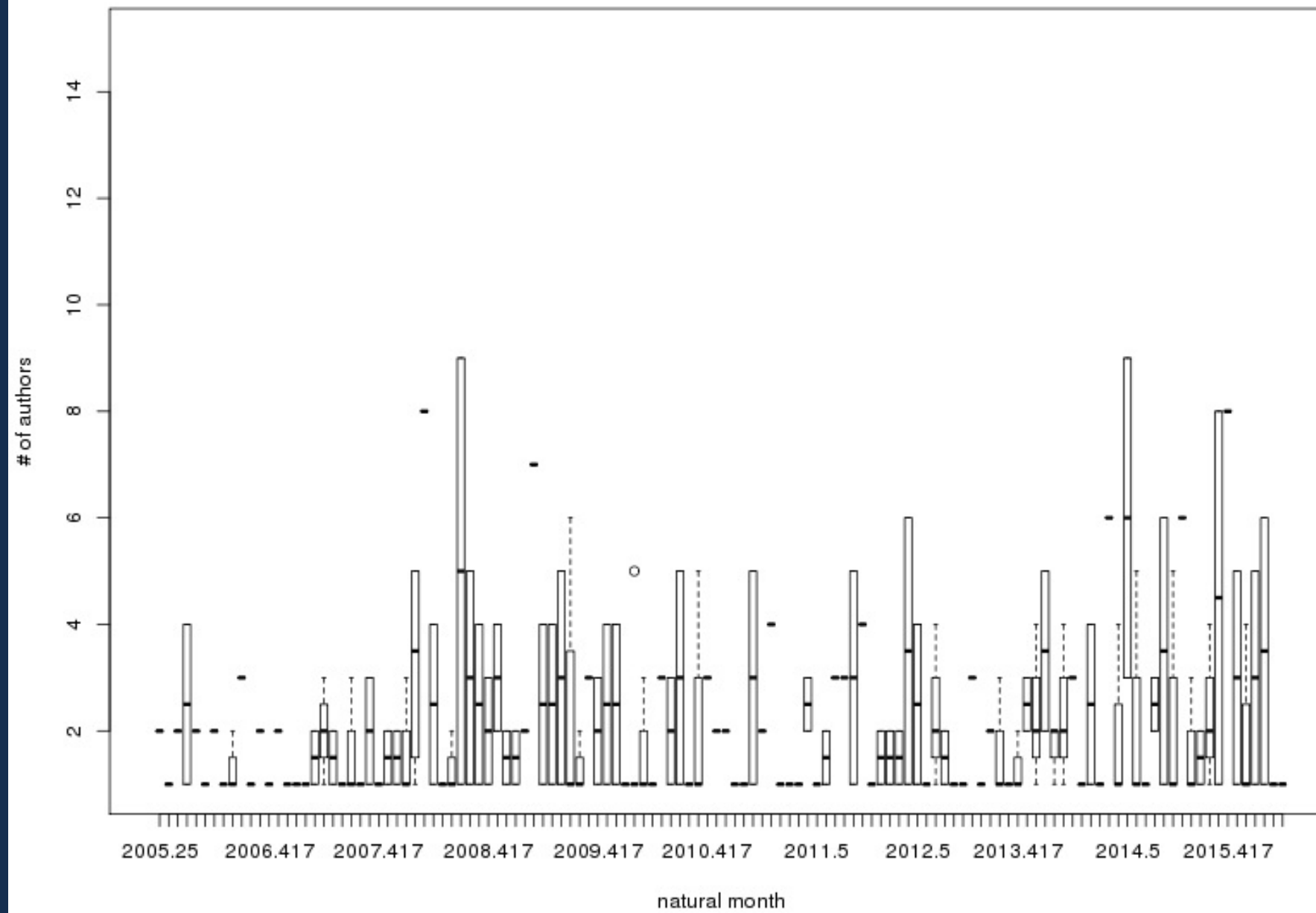
- linux kernel贡献难度的分析
- linux kernel performance test

个人进展

- linux kernel贡献难度的分析
 - 参考了北大师生的相关工作
 - 收获：数据的预处理模式值得参考
 - 重现
 - 结果分析展示



Boxplot of # authors per committer committed code for in crypto



个人进展

- linux kernel performance test
 - 目的：研究对于linux kernel而言最容易引起PB的是哪种benchmark、哪一类module、哪些基础设定
 - 使用了intel 提供的lkp-tests数据（60G）
 - 数据抽取、清洗、聚类
 - 目的：方便、准确处理，节约分析成本
 - 主要内容
 - 解析、抽取数据内容
 - 分辨、剔除不完整数据
 - 数据整合，补全
 - 结果 归一化的数据

个人进展

- linux kernel performance tests
 - 降维（主成分分析）
 - 目的：凸显主要的影响因素、降低分析成本（数据量）
 - 主要内容：
 - 估计主成分个数
 - 提取主成分
 - 主成分意义的分析和解释

主成分意义的分析和解释

PC3	
aim7.time.page_size	0.81
time.page_size	0.81
time.page_size	-0.17
time.elapsed_time	-0.50
aim7.real	0.32
time.involuntary_context_switches	0.13
aim7.time.elapsed_time	0.32
time.maximum_resident_set_size	0.06
aim7.time.percent_of_cpu_this_job_got	0.52
time.system_time	0.15
aim7.jti	0.80
aim7.jobs.per.min.per.task	0.73
aim7.time.maximum_resident_set_size	0.57
aim7.cpu	0.15
aim7.jobs.per.min	0.69

结果

- dimensions 2000+ -> 27
- 方差保留率90%+

lkp-compare的初步结果举例

主成分与benchmark的相关度研究

netperf	0.76
autotest	0.73
piglit	0.64
trinity	0.48
ebizzy	0.45

(共性分析尚未完成)

后期的安排

通过benchmark和性能数据对linux kernel的性能特征和变化情况进行分析

- 对比和分析三种性能缺陷检测方法的异同、优缺点（一到两周）
- 基于以上三种方法（或改进方法）量化性能缺陷对linux kernel的性能总体影响（一到两周）

谢谢！