Cooperation and Social Status in Free and Open Source Projects

Gaudenz Steinlin
gaudenz@debian.org

DebConf 9, Caceres, Spain

Introduction

- About my masters thesis in sociology
- Based on empirical process data gathered from the Debian project
- Better title: Size and Evolution of Debian
- Focus on what's interesting to Debian contributors

Presentation available at
`http://people.debian.org/~gaudenz/`

Outline

Data Sources

- Archives of all lists.debian.org mailinglists
- Package uploads (from debian-devel-changes and related mailinglists)
- Bug logs from bugs.debian.org
- Debian keyring changelog
- Popcon statistics (not used)
- ~38 GB of data

The dataset contains everything from April 1995 up to November 2007. 38 GB of data

Feeding the Database

1. Download raw data from Debian Servers
2. Data analysis with data specific Python scripts
3. Results stored in one big relational database (PostgreSQL)

Collected data

- Mails
  - Person
  - Time and timezone
  - List or Bug number
  - Words
  - References
- Uploads
  - Uploader and signer
  - Other contributors
  - Number of changelog entries per person
  - Package name
- Bugs
  - Submitter and fixer
  - Package
  - Severity
  - Way of closeing
  - Mailcount

Cleaning the data

Problems:

- Duplicated entries for the same person
- Role addresses
- Unparseable mails (mostly spam)

Deduplication:

- GPG Key Uids
- Semi automated using FEBRL `http://datamining.anu.edu.au/projects/linkage.html`
- Based on realname, nicknames, message-ids, domainnames, lists, packages
- A lot of manual work (more than a month)
- Only done for bug submitters and developers
- People only writing to list were removed from the dataset
- Impossible to manually deduplicate list contributors

Cleaning the data

Problems:

- Duplicated entries for the same person
- Role addresses
- Unparseable mails (mostly spam)

Deduplication:

- GPG Key Uids
- Semi automated using FEBRL `http://datamining.anu.edu.au/projects/linkage.html`
- Based on realname, nicknames, message-ids, domainnames, lists, packages
- A lot of manual work (more than a month)
- Only done for bug submitters and developers
- People only writing to list were removed from the dataset
- Impossible to manually deduplicate list contributors

Aggregating the data

- Data summarized by month and person
- Processed with a Python script:
  - Running on 40 nodes of a computer cluster for about 20 hours
- Second dataset about bug reports
  - Combines bug reports information with personal attributes of submitter and fixer
  - At the time the bug was submitted
  - Status of submitter and fixer
  - Interaction between submitter and fixer

Final dataset structure

- person
- start_period
- end_period
- last_update
- bugs_submitted
- bugs_fixed
- bug_activity_mails
- bug_activity_words
- mails
- mail_words
- uploads
- uploads_signed
- package_work
- patches
- core_mails
- core_words
- packages
- bugs
- bug_cooperations
- package_cooperations

Status of Debian Contributors

Contributor At least one entry on bugs.debian.org. Mostly bug submitters, mostly unknown to other project members

“Simple” developer At least one contribution recorded in a package changelog. Probably known to other team members.

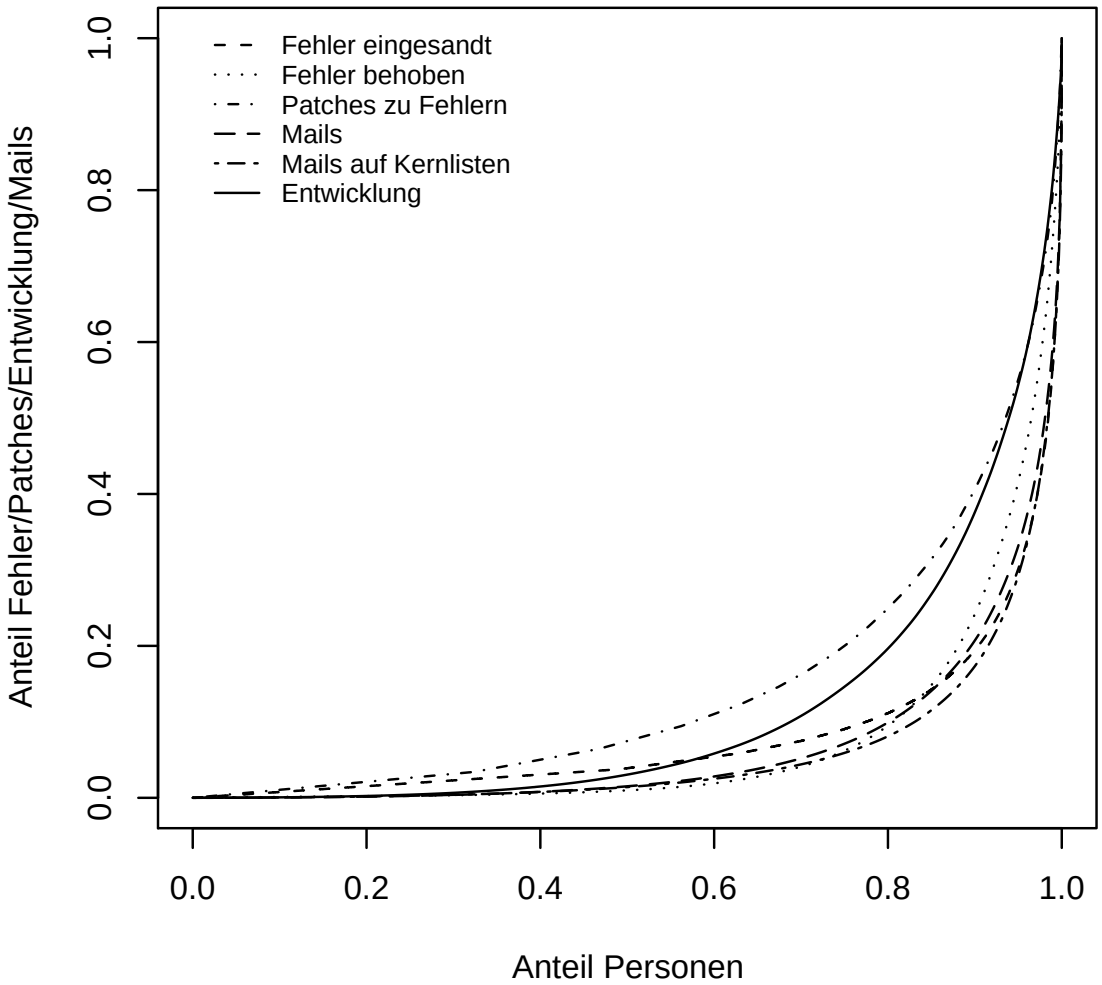
Official developers Can sign and upload packages. Most interaction partners will recognize official developers.

Core developers Those 20% of developers that have contributed the most to Debian (package work). Together they do about 80% of the work. Most are DDs. They are well known among other project contributors.

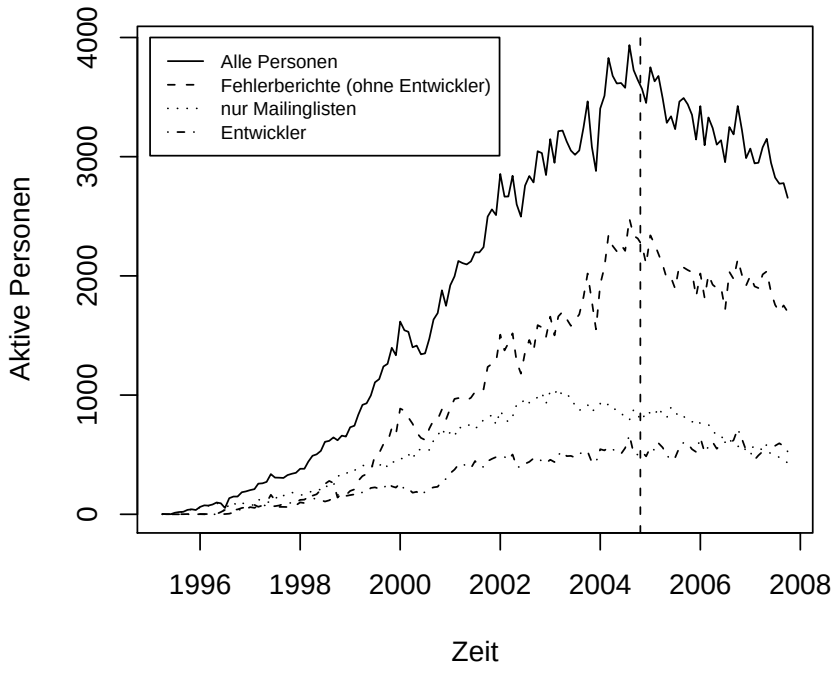
How many people are involved in Debian?

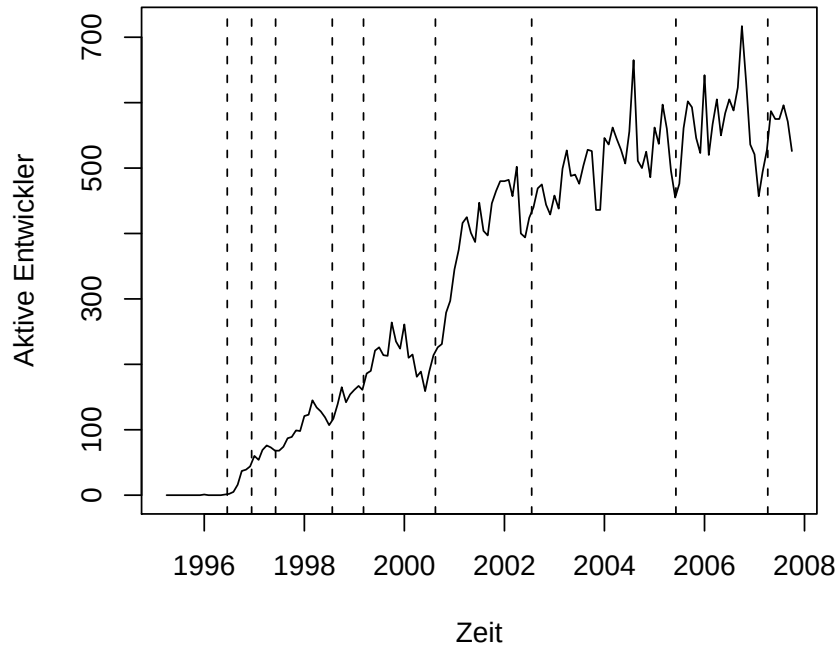
	N	(active)			
Whole dataset	236620		100%		
Deduplicated dataset	32538	(7206)	14%	100%	
Contributors	30027	(5647)	13%	92%	
Developers	2512	(1559)	1%	8%	100%
“Simple” developers	1181	(745)	<1%	4%	46%
Official developers	827	(390)	<1%	3%	35%
Core developers	503	(424)	<1%	2%	20%

Distribution of work in Debian

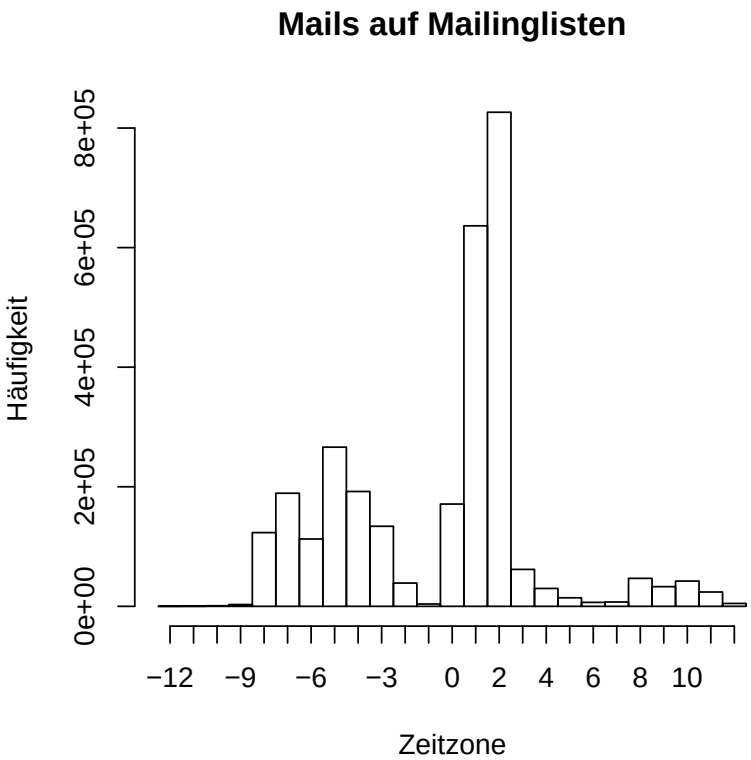


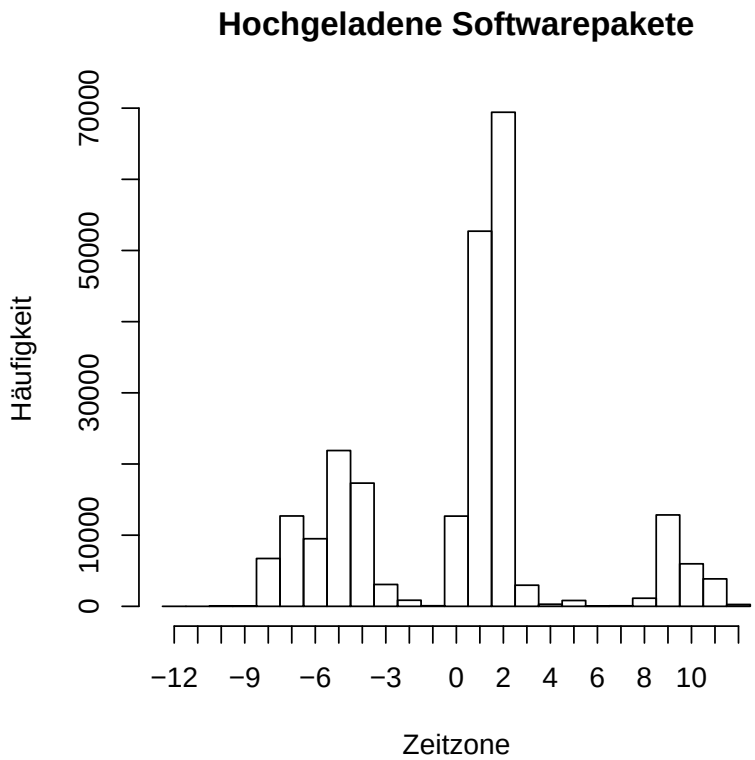
Temporal evolution of work done on Debian





Regional Distribution of Developers (1)

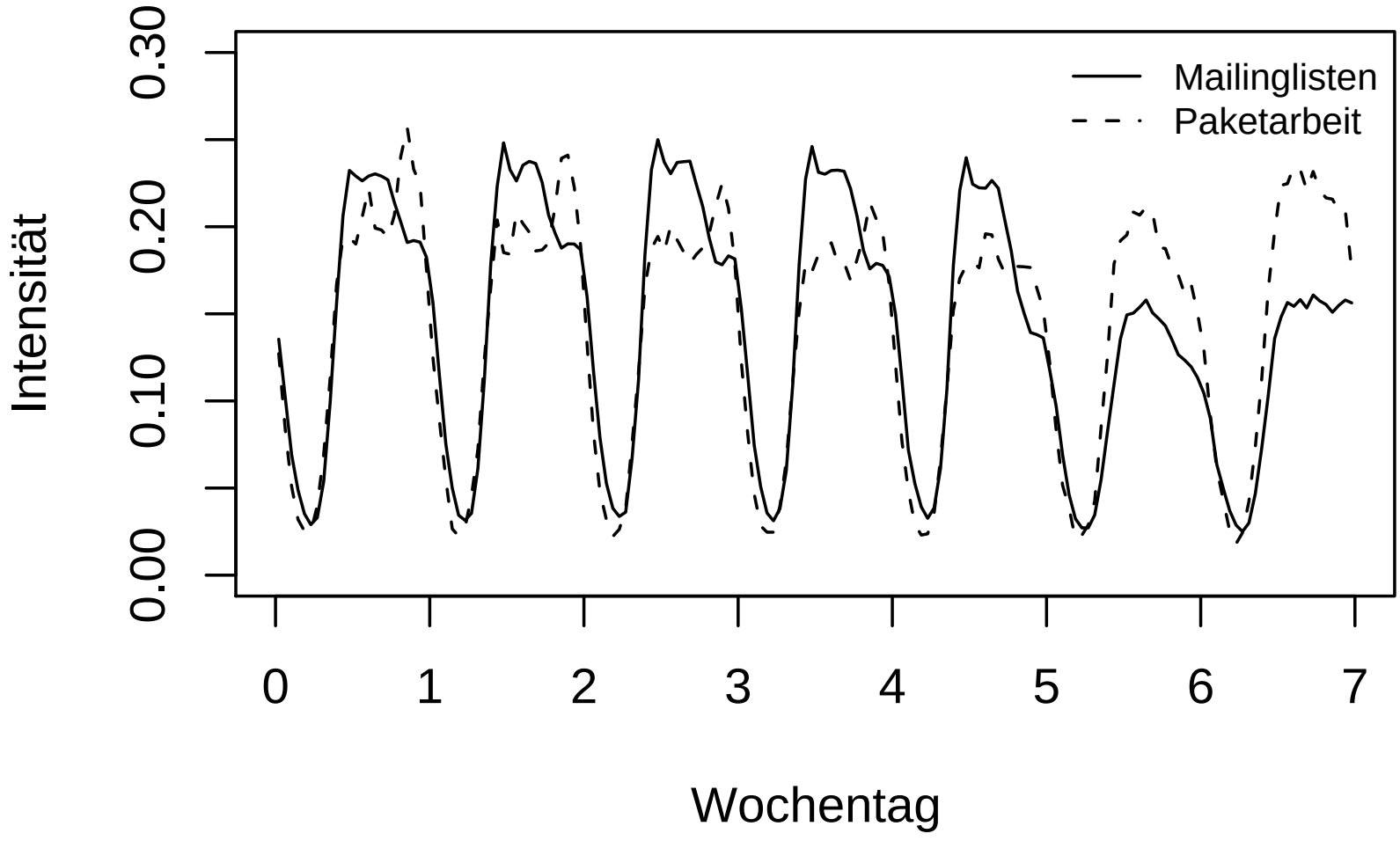




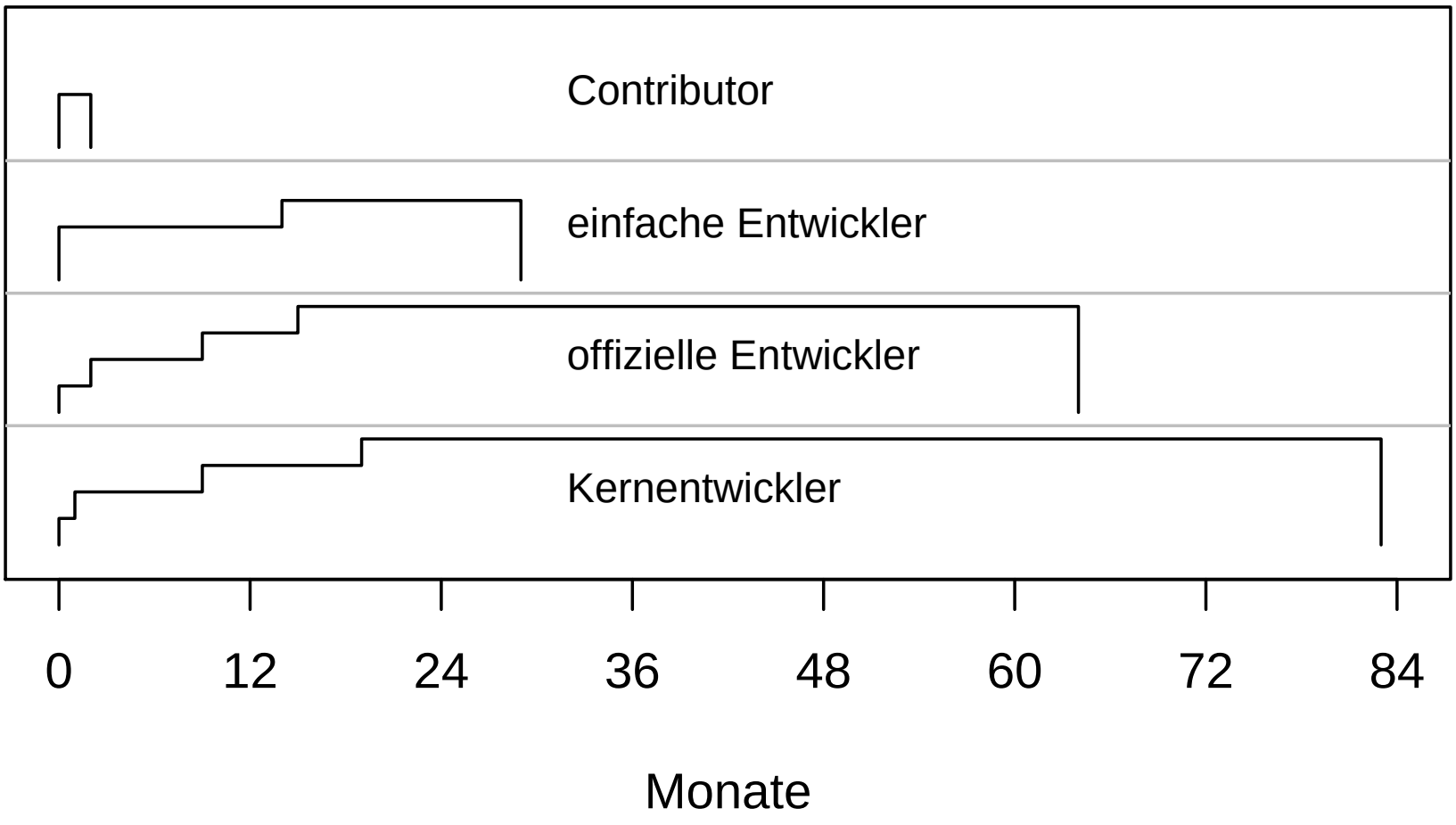
Regional Distribution of Developers (2)

Region	TZ	Lists	Packages	POLS	US
America	-8 – -2	34%	30%	14%	27%
Western Europe	0 – 2	55%	57%	70%	53%
Australia & Japan	8 – 11	5%	10%		

Temporal Distribution of Work



Developer “carriers”



Active and inactive Periods

	Num. Periods	Median Length		% of Total	
		Active	Inactive	Total	only Dev.
All Persons	3.57	1	2	65%	3%
Contributors	3.24	1	3	65%	–
“Simple” Devels	6.38	2	2	58%	34%
Official Devels	10.16	2	2	60%	44%
Core Devels	6.27	3	1	86%	70%

Prisoners dilemma

	Cooperation	Defection
Cooperation	3,3	0,5
Defection	5,0	1,1

- Rational actors won't cooperate
  - Higher payoff of defection regardless of the others choice
- Both would be better off if they would cooperate

Iterated prisoners dilemma

- Iteration (playing again) can lead to cooperation
  - Tit for tat strategy: Players start with cooperation and do what the other did in the last round
  - Instant punishment
  - Can forgive
- Does not scale well to FOSS developement: Only works if developers interact several times
- Extensions to N-player games exist, but cooperation is very fragile

How to stabilize cooperation?

- Developers need additional information about their interaction partners
- Reputation
  - Not directly observable
  - Assumption: Developers build up reputation by contributing to the project
  - Work done for the project is a proxy for reputation

Reputation in Free Software Developement

“The “utility function” Linux hackers are maximizing is not classically economic, but is the intangible of their own ego satisfaction and reputation among other hackers.”

Eric Raymond

*“[...] statements from people who *do* things always carry more weight with me than people whose primary activity is making statements.”*

Steve Langasek on debian-vote

Reputation in Free Software Developement

“The “utility function” Linux hackers are maximizing is not classically economic, but is the intangible of their own ego satisfaction and reputation among other hackers.”

Eric Raymond

*“[...] statements from people who *do* things always carry more weight with me than people whose primary activity is making statements.”*

Steve Langasek on debian-vote

“A reputation for honesty, or thrustworthiness, is usually acquired gradually. This alone suggests that the language of probabilities is the right one in which to discuss reputation: a person’s reputation is the ‘public’s’ imputation of a probability distribution over the various types of person that the person in question can be in principle.”

Partha Dasgupta, Economist

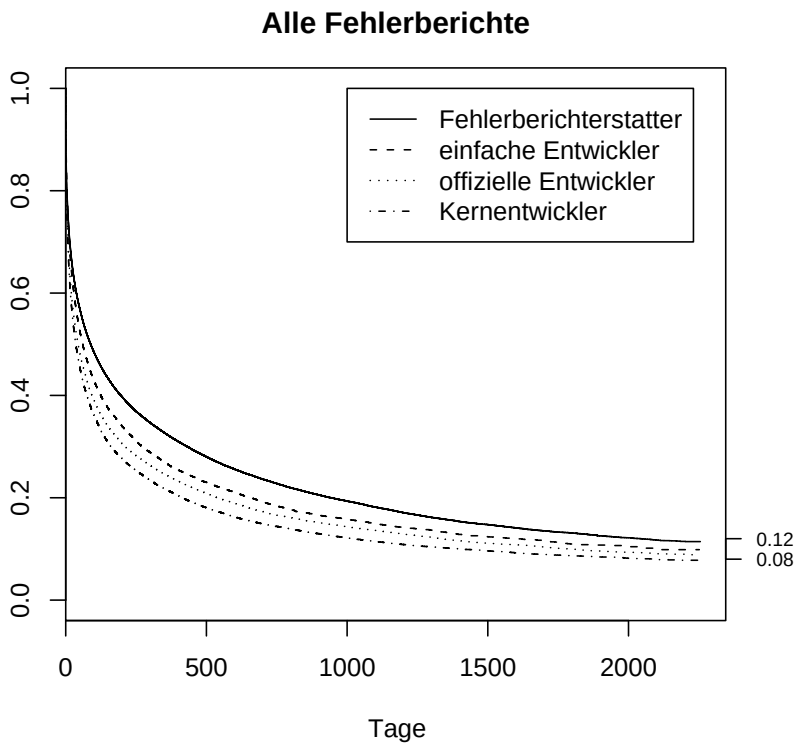
Empirical verification

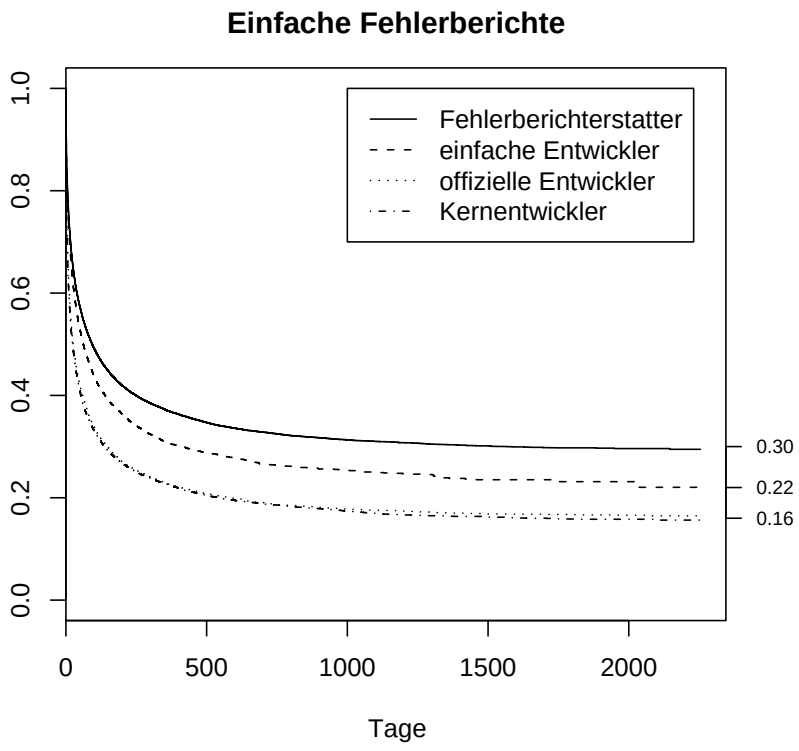
- Does reputation have an influence on bug fixing?
- Hypothesis: The higher the status of the bug submitter the faster a bug will get fixed.

Bugfix time depending on Developer Reputation

Status	N	Events	Median	95% conf.int.
Complete Dataset				
All	305'342	241'661	63.4	62.5 – 64.3
Contributors	170'799	132'378	88.6	86.9 – 90.4
"Simple" Devs	31'923	23'702	60.5	58.1 – 62.6
Official Devs	33'841	28'494	44.0	42.3 – 45.7
Core Devs	67'231	55'844	35.8	34.9 – 36.8
Reduced Dataset				
All	131'382	89'917	53.8	52.8 – 55.1
Contributors	67'496	43'041	93.2	90.6 – 96.4
"Simple" Devs	15'824	9'935	61.8	58.8 – 65.9
Official Devs	15'132	11'822	26.3	24.5 – 28.1
Core Devs	32'314	24'703	23.7	22.6 – 24.7

Survival Curves of Bugs

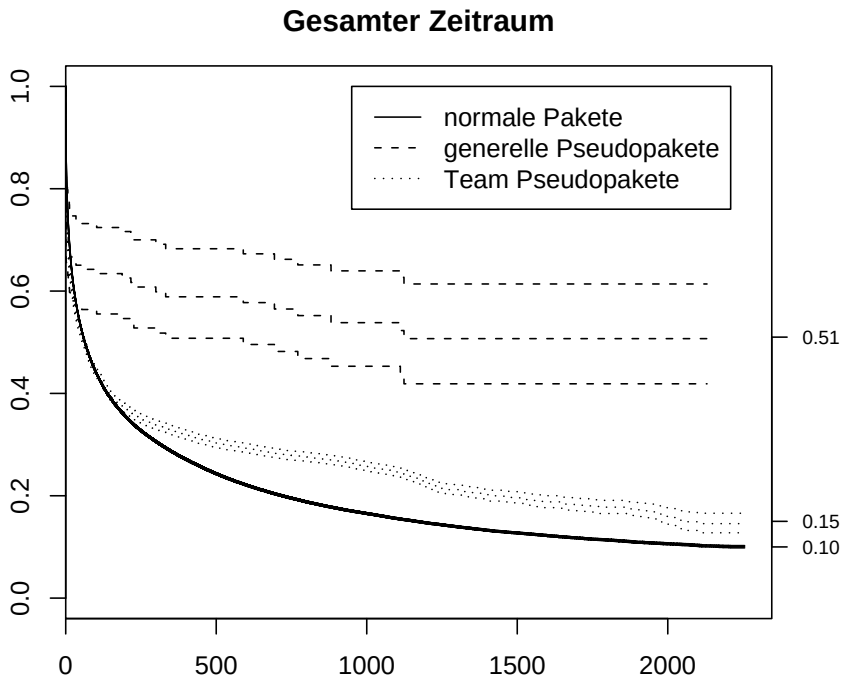


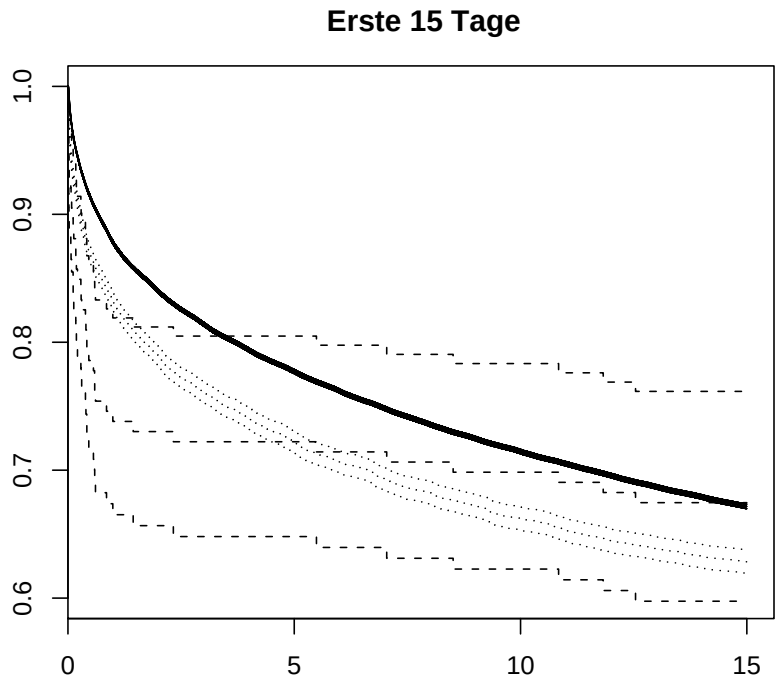


Results from multivariate survivial analysis

- Control variables: bug severity, bug fixer, experience of submitter, pseudopackages, way of fixing, ...
- Significant results for most reputation variables
- Prior interaction has a strong influence
- More significant and higher effects in the reduced dataset with simple bugs
- Relatively low explanation of overall variance

Pseudopackages





Outlook

- Much more could be done with this data
  - Network analysis
  - Analysis including mailinglists
  - ...
- Data is available only, bigger files only on request

Further References

- Thanks for listening!
- Full Text in German, All scripts I used:
  `http://gaudenz.durcheinandertal.ch/thesis/`
- Printed version of the thesis available
- Any questions or comments?