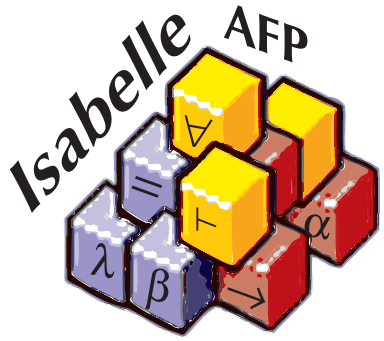
Mining the Archive of Formal Proofs

Jasmin Blanchette Max Haslbeck
Dan Matichuk Tobias Nipkow



What does an archive of machine-checked proofs look like today?

What are the processes?

Can we quantify the contents?

How well do ATPs perform on it?

This talk: A snapshot of one such archive

1. Introducing the AFP
2. Sizes and Structure
3. Proof Automation
4. Conclusions

The Archive of Formal Proofs (AFP):

- an online library of proofs
- for the proof assistant Isabelle
- contributed by its users

- Each contribution is a collection of theories
- Each theory is a sequence of
  definitions, lemmas, proofs

- Each submission is reviewed (lightly)
  by an editor (Klein, Nipkow, Paulson, Thiemann)

A guided tour of the AFP

http://isa-afp.org

Start of AFP: 2004.

July 2017:

362	articles
97,000	lemmas (= theorems)
1,700,000	lines

Related archive: Mizar Mathematical Library

- Start: 1989
- About 3 million lines, 50,000 lemmas
- Heavily biased towards mathematics
- Maintained and continuously revised by Mizar developers
- Logic: set theory

Related archive: Coq-Contrib GitHub Repository & Coq OPAM packages

- Start: 1992?
- About 1.5 million lines, 54,000 lemmas, 168 entries
- Computer science and mathematics
- Maintained by Coq developers
- + 72 user-maintained OPAM packages (e.g. MathComp)
- Logic: type theory

Characteristics of the AFP

- Often proofs for published articles
- More applications than foundations
- Structure and coherence not imposed by AFP, must come from authors
- Biased towards computer science
- Logic: HOL

Below the AFP

Archive of Formal Proofs

Isabelle/HOL

- Comes with many basic libraries:
  sets, relations, lists, algebra, number theory, analysis, probability theory, ...
- 500,000 lines
- Maintained and continuously revised by the developers

AFP maintenance

- Isabelle and AFP releases in sync
- Release version of AFP is frozen
- Development version of AFP is updated:
  - by Isabelle developers to keep it in sync with Isabelle
  - by AFP authors to improve/extend articles
- Each change to the Isabelle or AFP repository triggers an incremental test run

The boon and bane of AFP maintenance

+ Release version:
  All AFP articles “work”
+ Development version:
  Most AFP articles work most of the time
− Significant overhead for developers
− Even “dead” articles are maintained
+/− Can make changes in Isabelle or its libraries too costly

Some Questions

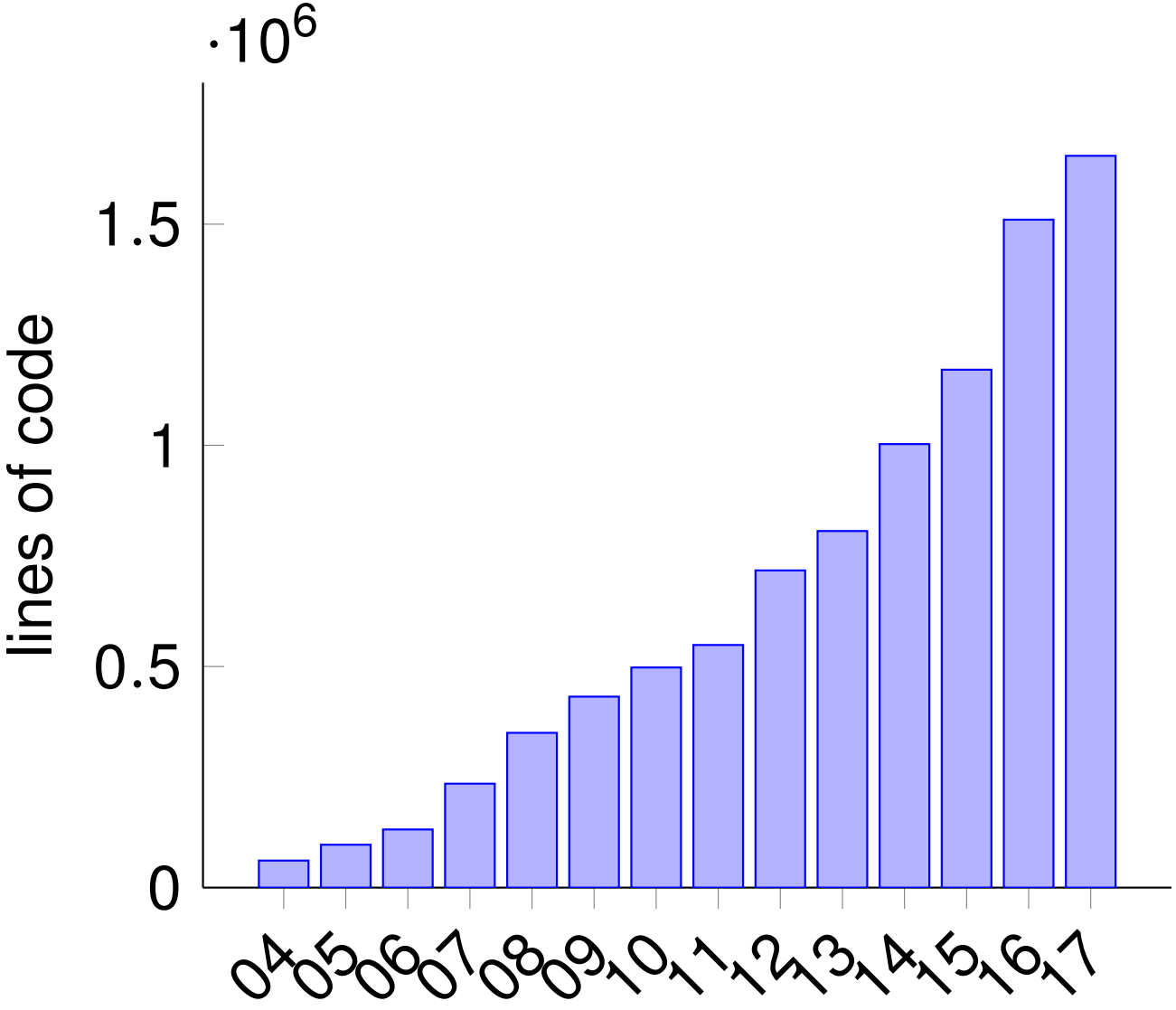
- How large are articles?
- Size of definitions vs. lemmas vs. proofs?
- How many contributors?
- How large are their contributions?
- How much reuse?
- Characteristics of the dependency graph?
- How did the AFP evolve over time?
- Can we estimate the size of a proof from the statement to be proved?
- How many equalities, Horn clauses, . . . ?
- Performance of Sledgehammer on AFP?

LOC = lines of “code”

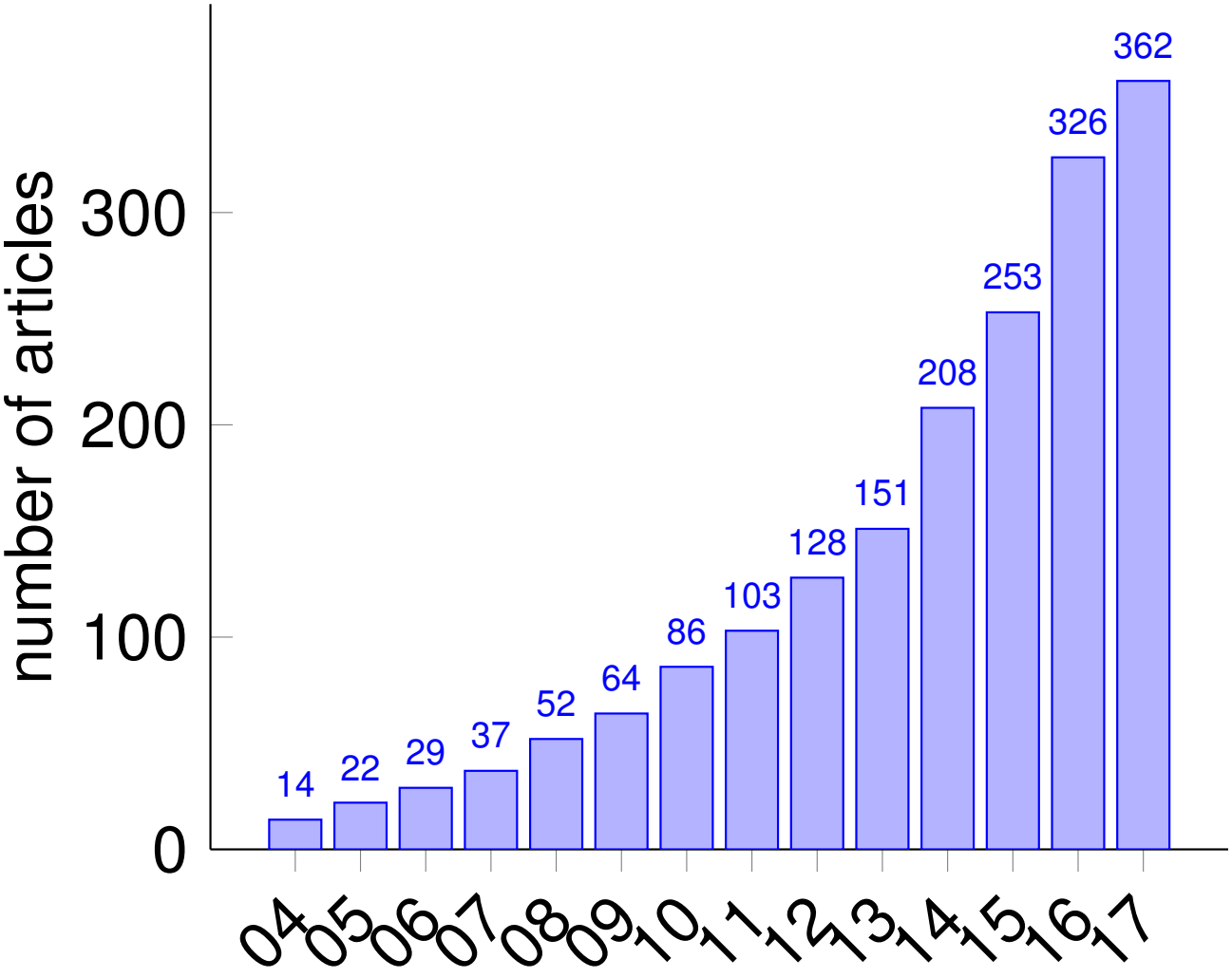
= lines of Isabelle text

Number of lines, articles, authors over time?

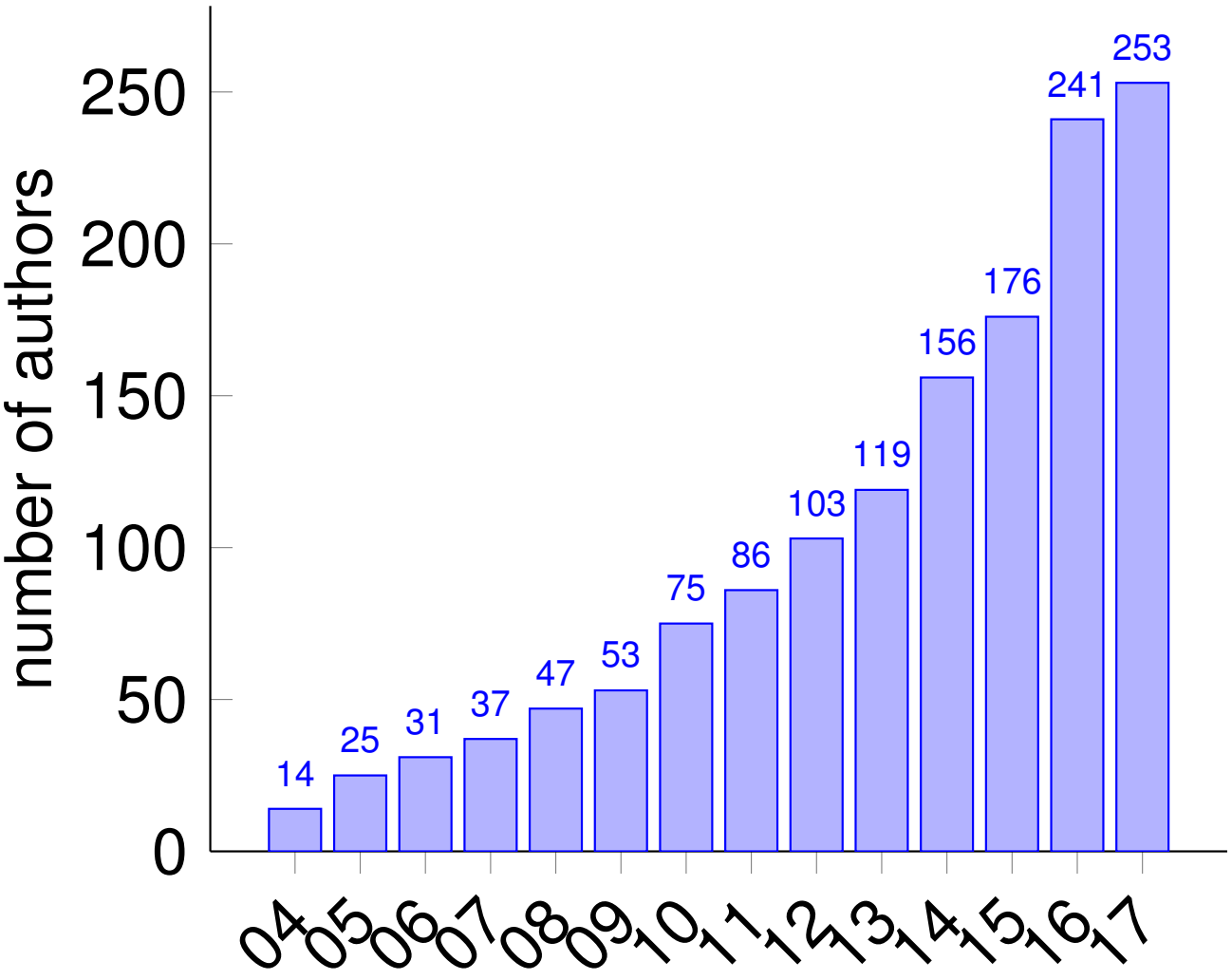
Number of lines (cumulative)



Number of articles (cumulative)



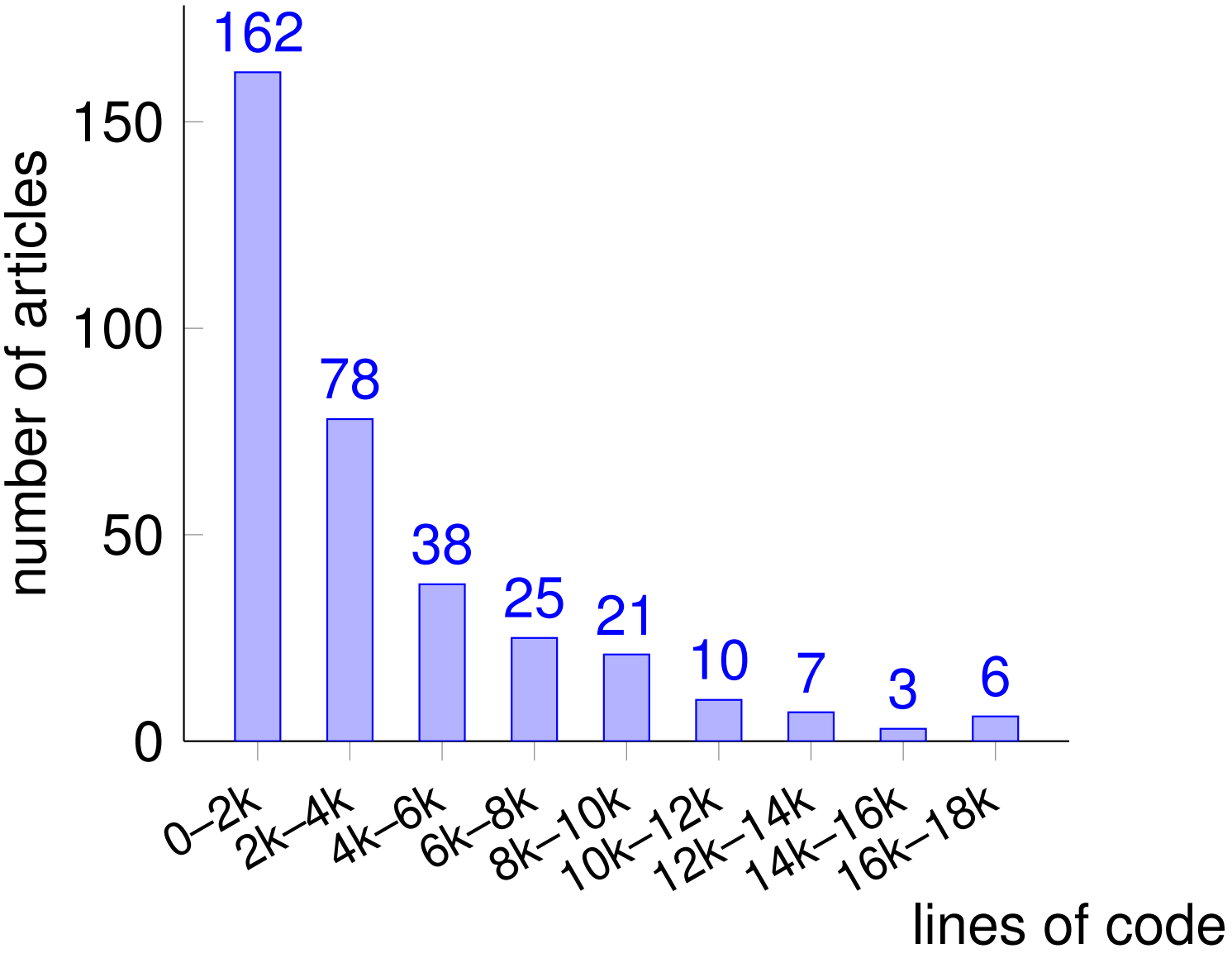
Number of authors (cumulative)



What is the average size of an article?

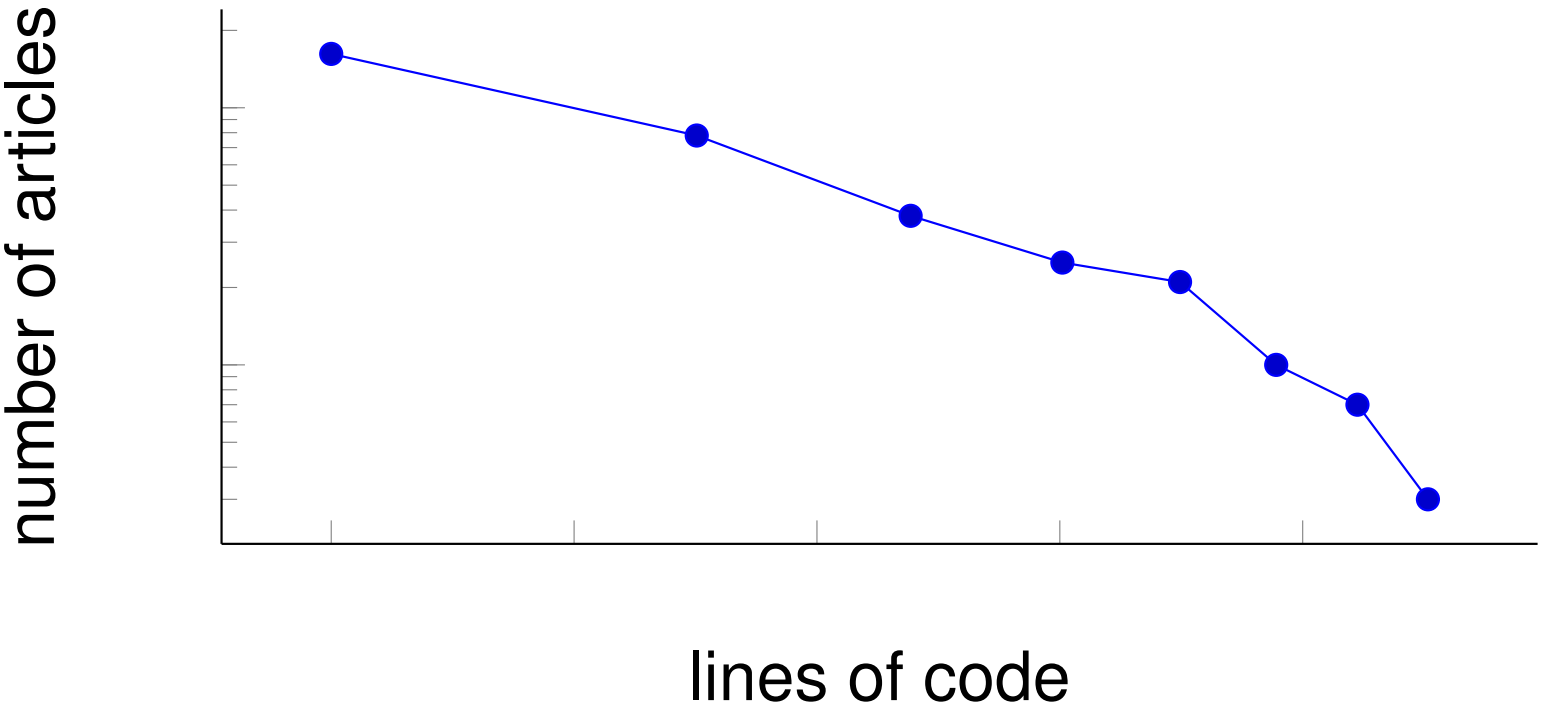
4,600 LOC

Sizes of articles



Distribution of article sizes?

log-log plot:



Straight line $\Longrightarrow$ power law

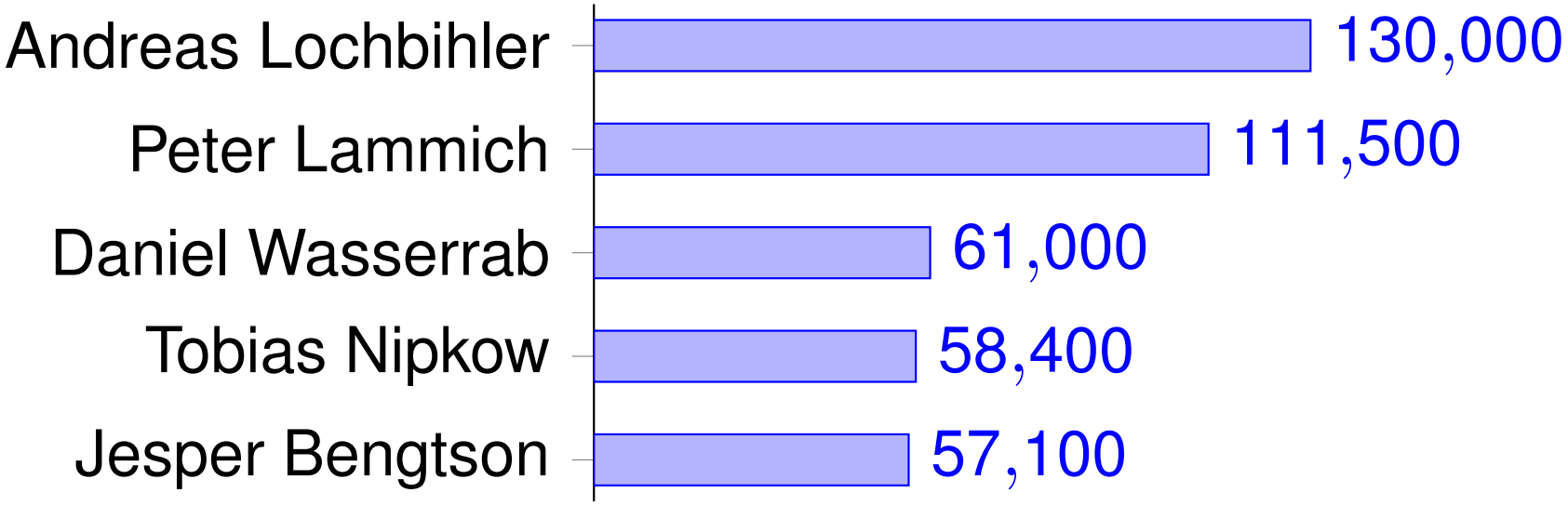
Power laws for many similar distributions, eg file sizes

Long tail

- 10 articles are $>$ 20,000 LOC
- Largest article:
  `JinjaThreads`: 77,000 LOC
  by Lochbihler

Who are the top contributors to the AFP?

Top 5 contributors by LOC



Topics: programming languages
program analysis
model checking
data structures
Flyspeck
combinatorics

Sizes of definitions vs. lemmas vs. proofs?

Definitions:	9%
Lemmas:	17%
Proofs:	60%
Other:	14%

Are these numbers typical?

Comparison of provers

Wiedijk (2007):

	AFP	Isa	Coq	HOL Light	Mizar
Definitions:	9	8	10	1	3
Lemmas:	17	21	12	14	9
Proofs:	60	50	60	62	84

Some interpretations:

- *Pure math*
- *Automation*

How many lemmas per definition?

Archive of Formal Proofs: 3.6
Odd Order Thm in Coq: 3.3

One good definition is worth three theorems.

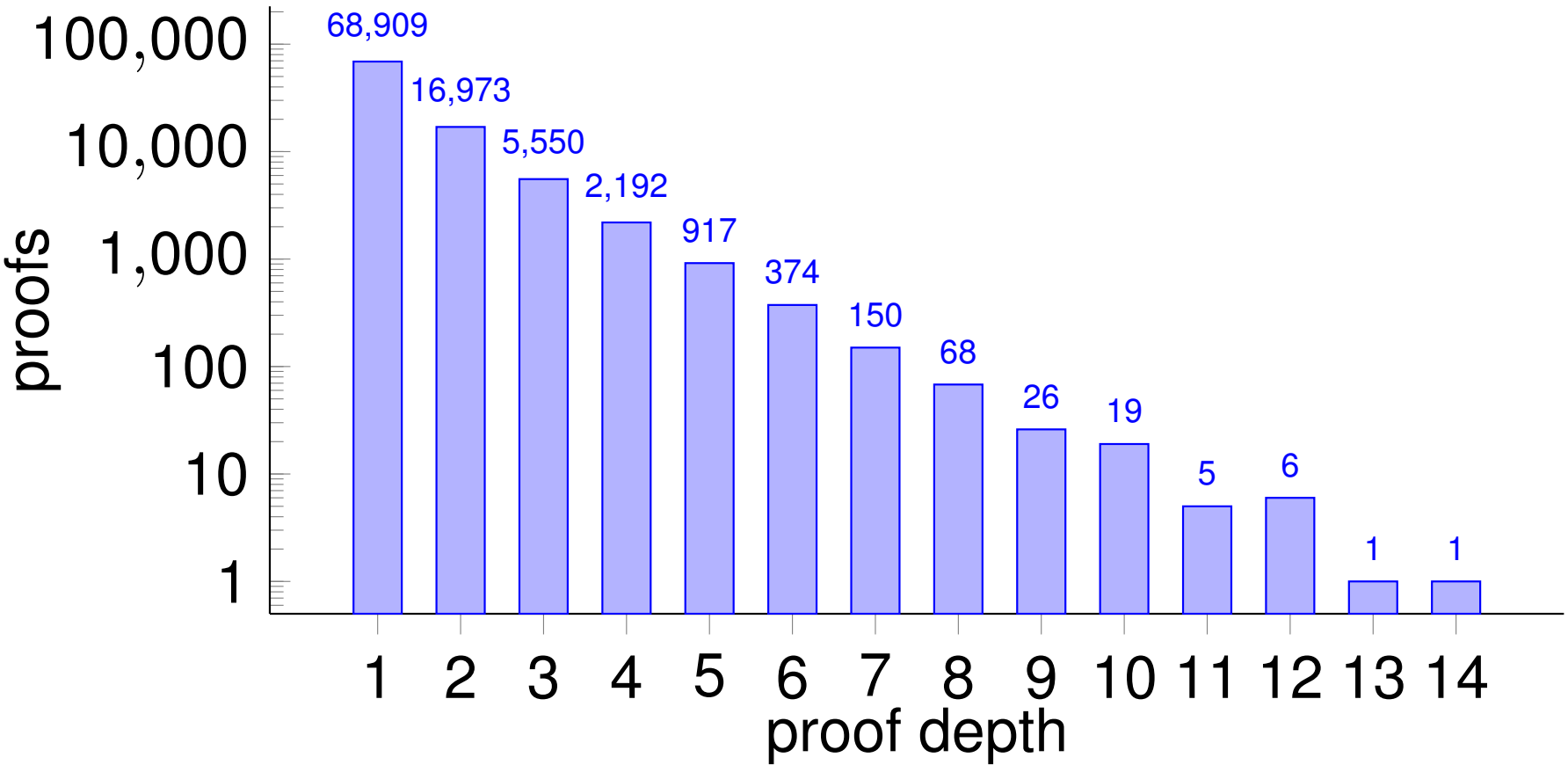
— Alfred Adler, "Mathematics and Creativity,"
The New Yorker (1972)

How deep is your proof?

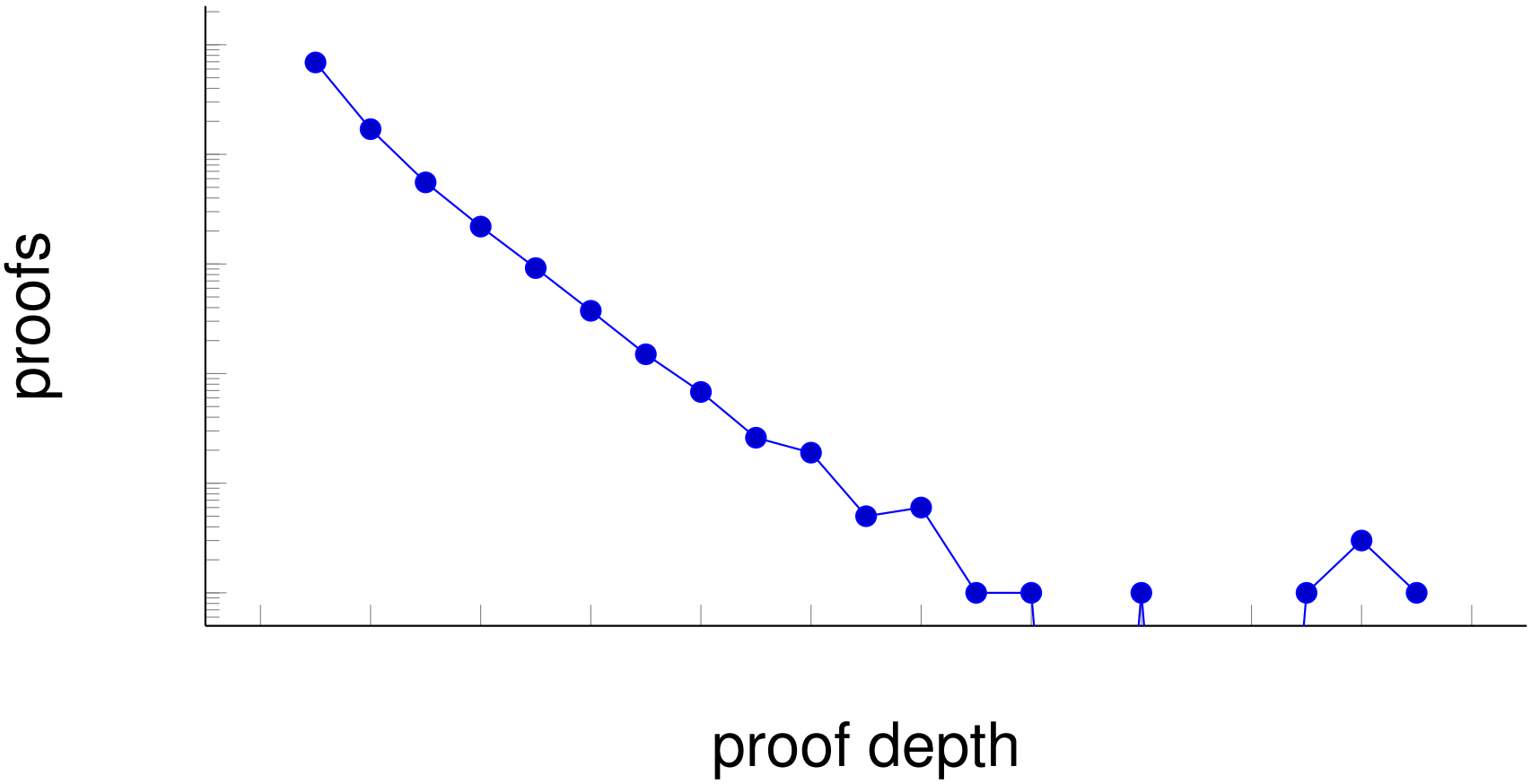
Proofs in Isabelle are block-structured

How deep are AFP proofs?

How deep is your proof?



How deep is your proof?

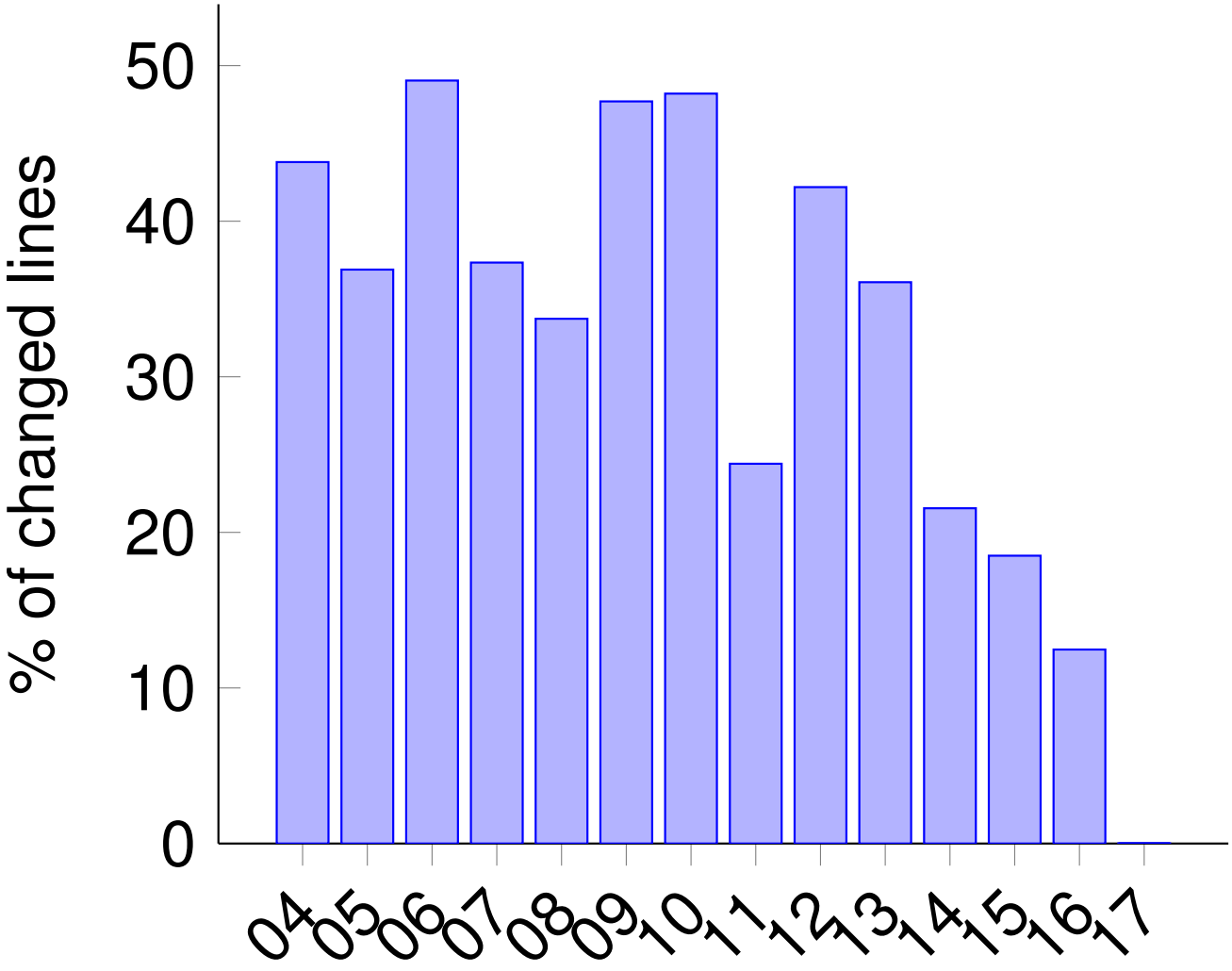


Straight line $\Longrightarrow$ exponential decay

How much do articles change over time?

On avg: 25%

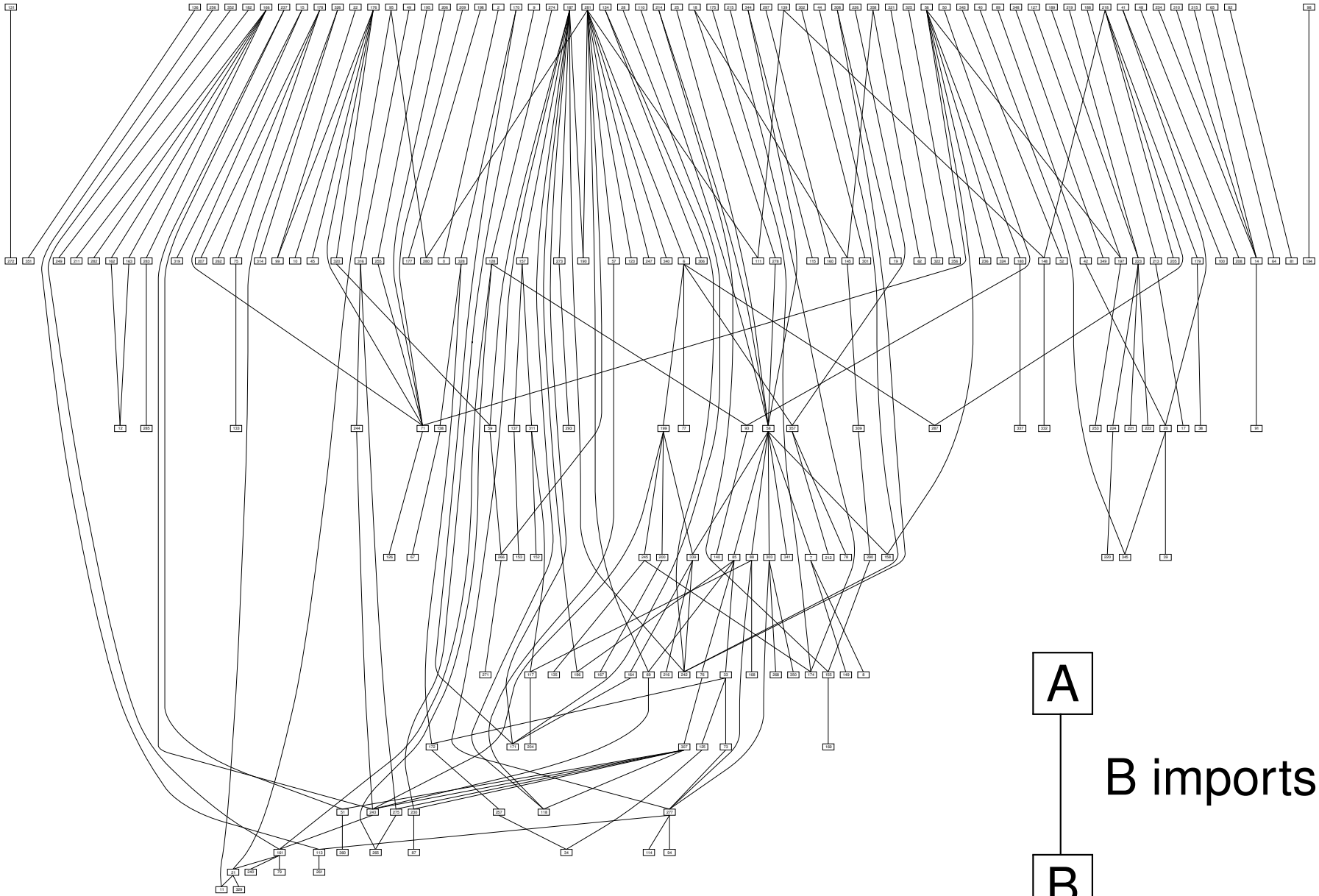
Changes over time



Percentage of changed lines for articles from year *x*

How much is actual development,
how much merely maintenance?

AFP import DAG

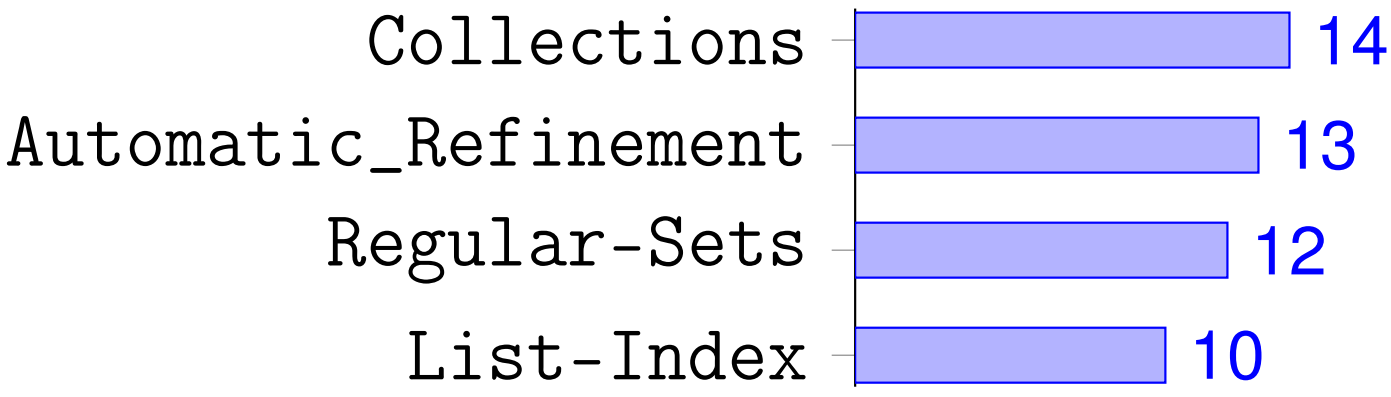


Graph metrics

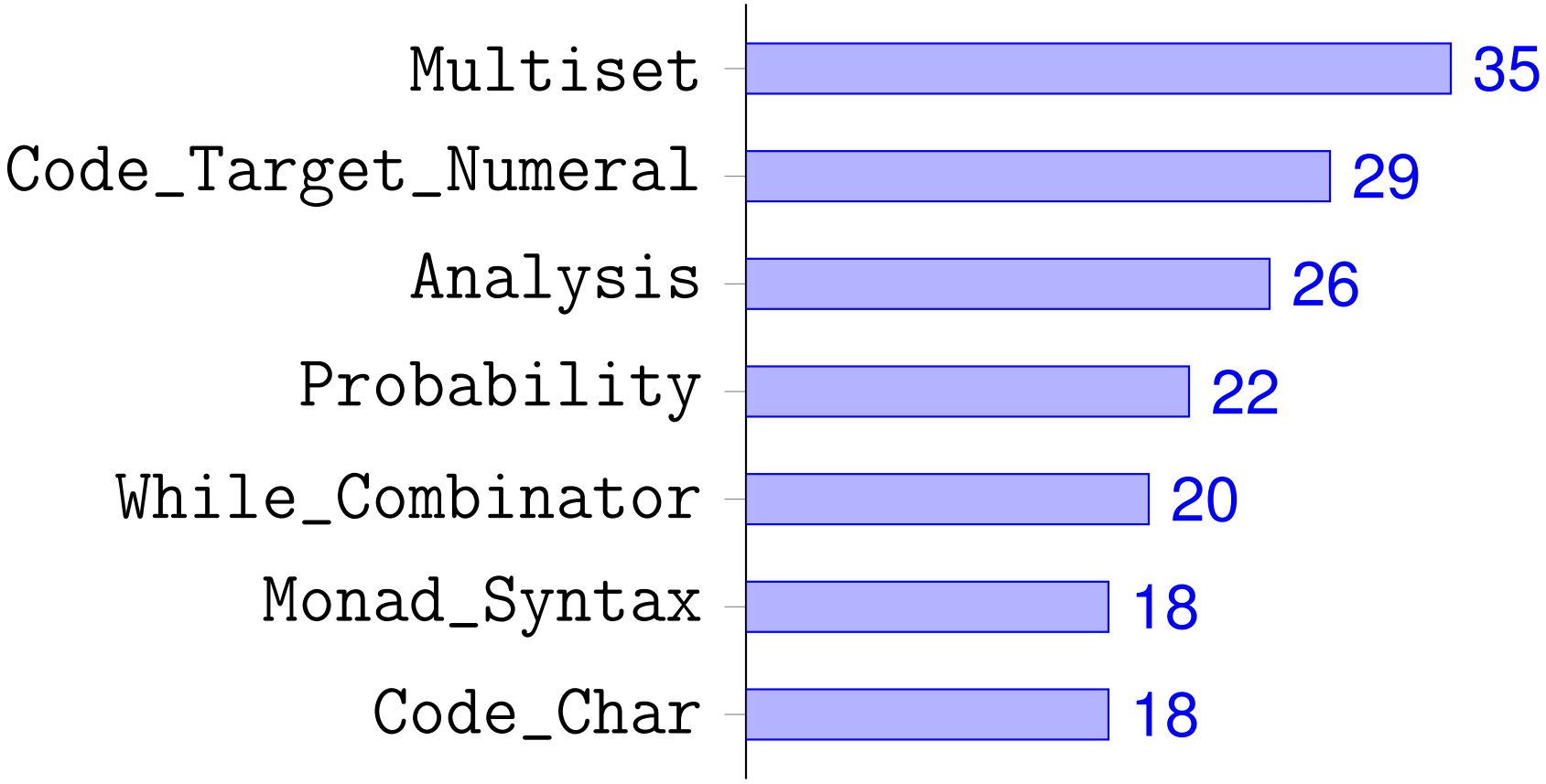
Transitive edges deleted

Nodes	362
Isolated nodes	146
Edges	217
Depth	9
Max. out-degree/export	10
Max. in-degree/import	6

Most popular AFP articles



Most popular library theories



Least popular AFP articles

237 of 362 articles are never used

How bad is that?

What percentage of journal articles is never cited?

"90% of papers in academic journals are never cited"

L.I. Meho: *The rise and rise of citation analysis.*

No data or source given

What do other sources say?

“50% of these publications had no citation in WoS”

“23.7% of articles had not been cited”

“3% of the papers are never cited”

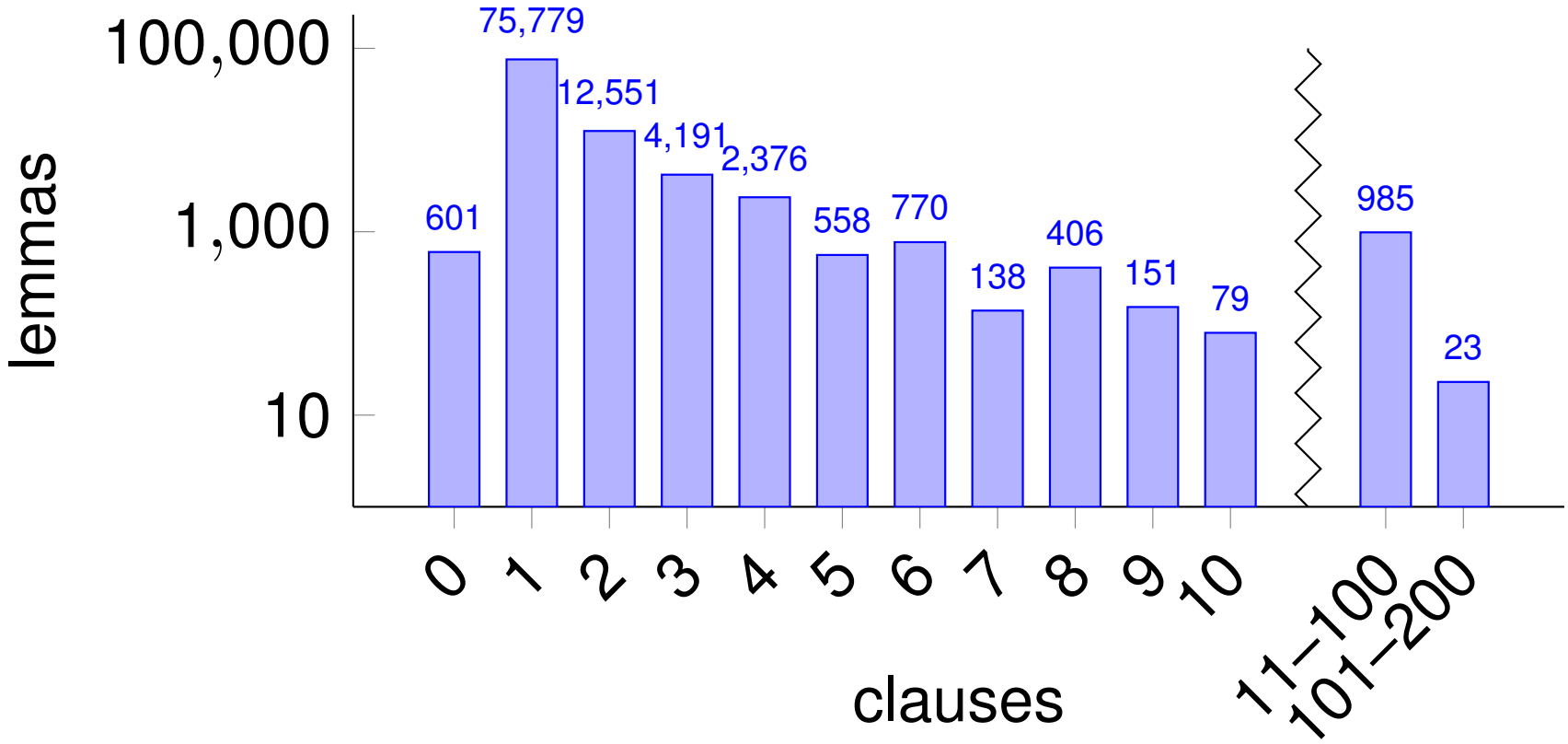
(J. Fluid Mechanics)

Complexity measure: CNF

Why CNF? Most automatic provers work with CNF.

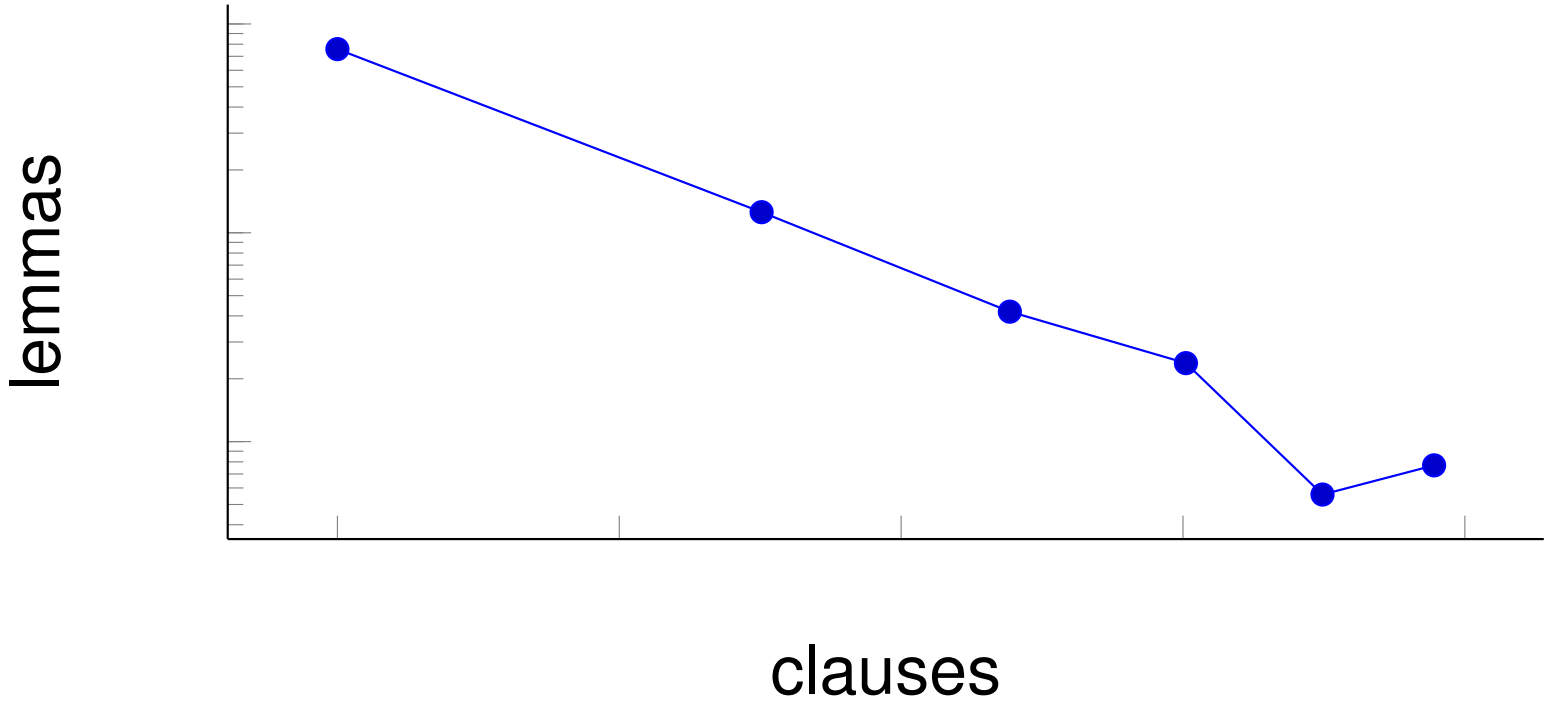
- CNF is a conjunction of clauses
- Clause is a disjunction of literals
- Literal is *atom* or $\neg$*atom*

Clause count



Clause count

log-log:



Power law

1 clause formulas: 73 %

Classification of 1 clause formulas:

Equations	23 %
Conditional Equations	46 %
Horn Clauses	92 %

67 % of all lemmas are Horn clauses

Sledgehammer (Blanchette & Paulson)

An Isabelle-interface to external ATPs:

1. Ships goal and (filtered) background library to external ATPs (CVC4, E, SPASS, Vampire, veriT, Z3, . . .)
2. Reinterprets external proof in Isabelle (can fail)

Nontrivial translations between logics!

Judgement Day

A benchmark suite of ~1250 goals from 7 “representative” theories (only some from AFP)

How many goals can be proved automatically within 30s:

2010: 46%

2015: 65% (without machine learning)
75% (with machine learning)

Reasons:

- Better translations
- Better provers, more provers
- Faster hardware

How representative is Judgement Day really?

New test data: A random AFP selection

- 128 random theories
- 100 random goals per theory (max)
- 8,921 goals overall

Evaluation (in %)

2017

	One-line	+ Isar	+ Oracle
CVC4	54.8	54.8	57.3
E	53.7	54.1	55.1
SPASS	53.1	53.4	54.3
Vampire	53.2	53.6	54.5
veriT	51.3	51.5	53.1
Z3	52.4	52.6	54.4

All 6 provers together with reconstruction:
63.3% (without machine learning)

Some conclusions

- More reuse!? (But AFP largely CS applications)
- AFP graph resembles citation graph
- Sizes follow power or exponential distributions
- 60% proofs, 17% lemmas, 9% definitions
- 34% (conditional) equations, 67% Horn clauses
- Sledgehammer can prove 65-75% of all goals

None of this is deep,
but it replaces anecdotal evidence
by empirical data.

(For details see CICM 2015)