2IS55 Software Evolution

Repository Mining: Social Aspects

Alexander Serebrenik

Technische Universiteit Eindhoven University of Technology

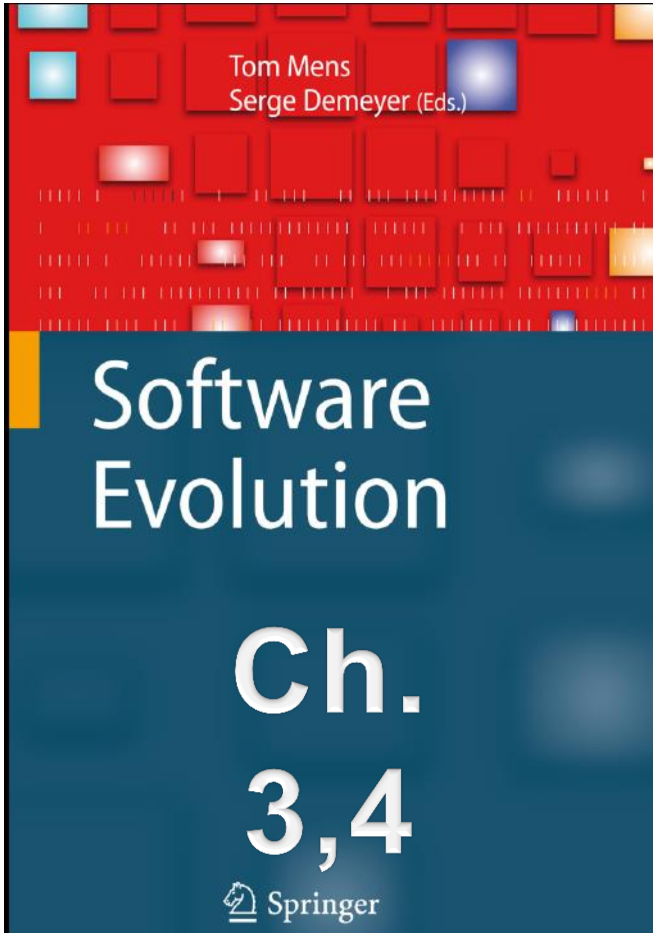
Where innovation starts

Assignment

- Assignment 2:
  - Deadline: Saturday

- Assignment 3:
  - Published on Peach
  - Deadline: March 17

Sources



Tom Mens
Serge Demeyer (Eds.)

Software Evolution

Ch. 3,4

Springer

Recap: Version control systems

- Centralized vs. distributed
- File versioning (CVS) vs. product versioning

- Record at least
  - File name, file/product version, time stamp, committer
  - Commit message

- What can we learn from this?
  - Humans TODAY !
  - Files
  - Bugs

Users in mail archives, version control systems, etc.

"On the Internet, nobody knows you're a dog."

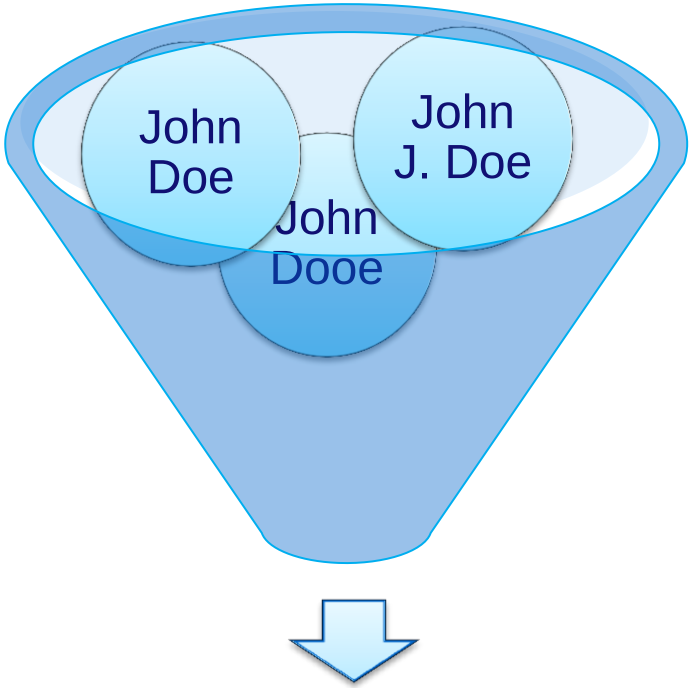
- Multiple aliases
  - a.serebrenik@tue.nl
  - aserebre@win.tue.nl
  - aserebrenik@yahoo.com
  - aserebrenik@gmail.com
  - alex@alum.cs.huji.ac.il
  - A.E.Serebrenik@cwi.nl

- Can be worse:
  - Ken Coar a.k.a. “Rodent of unusual size”
  - Aaron Brown a.k.a. Mrhappypants
  - KoffieTisch

What we want and what we need

- We would like to
  - Evaluate expertise
  - Evaluate contribution / involvement
  - Understand communication patterns
  - Study structure of the community (gender, country, education level…)

- We need to merge the aliases



id = 17, {John J. Doe, John Doe, John Dooe}

Identity merging

- **Input:**
  - **List of name, email address pairs**

- **Algorithms:**
  - **Simple: identical names, e-mail prefixes or user names**
  - **Bird: normalize names and cluster based on the Levenshtein distance** [Bird,Gourley,Devanbu,Gertz, Swaminathan 2006]
  - **LSA: combine the Levenshtein distance with latent-semantic indexing** [Kouters, Vasilescu, Serebrenik, van den Brand 2012]

Bird’s algorithm (1)

- Normalize names:
  - Remove punctuation and suffixes (“jr.”), reduce spaces and drop generic terms (“admin”, “support”)
  - Separate first name and last name

S	a	t	u	r	d	a	y
S	a	t	u	n	d	a	y
	S	a	u	n	d	a	y
		S	u	n	d	a	y

3 similarity measures

- Similarity of names
  - Levenshtein distance
  - Number of characters added, removed or modified
  - Names are similar if
    - either the full names are similar
    - or both the first and last names are similar

Bird’s algorithm (2)

- Similarity of names and mails
  - The prefix (before @)
    - Contains the first and the last names
    - Robles: Contains the first or the last name and the first letter of the other one
- Similarity of mails
  - Levenshtein distance on prefixes
- Cumulative similarity – maximal of the three

- Clustering based on the cumulative similarity
  - Large clusters
  - Human inspection and post-processing
    - It is easier for humans to split large clusters than to combine small ones

Still an heuristics!

How to calculate the Levenshtein distance?

- **Words X (n characters), Y (m characters)**
- **Data structure C[0..n,0..m]**
- **Init: C[i,0]=i, C[0,j]=j for any i and j**

C		S	a	t	u	r	d	a	y
	0	1	2	3	4	5	6	7	8
S	1								
u	2								
n	3								
d	4								
a	5								
y	6								

Similar to the longest common sequence (diff)

Technische Universiteit
Eindhoven
University of Technology

How to calculate the Levenshtein distance?

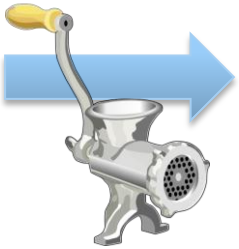
- **For every i and every j**
  - **If X[i]=Y[j] then C[i,j]=C[i-1,j-1]**
  - **Else C[i,j]=min(C[i-1,j]+1, // deletion**
    C[i,j-1]+1, // insertion
    C[i-1,j-1]+1) // modification

C		S	a	t	u	r	d	a	y
	0	1	2	3	4	5	6	7	8
S	1	0	1	2	3	4	5	6	7
u	2	1	1	2	2	3	4	5	6
n	3	2	2	2	3	3	4	5	6
d	4	3	3	3	3	4	3	4	5
a	5	4	3	4	4	4	4	3	4
y	6	5	4	4	5	5	5	4	3

The Levenshtein distance!

Algorithm of Kouters et al.

<John Doe, johnd@domainA>
<John Joseph Doe, johnd@domainA>



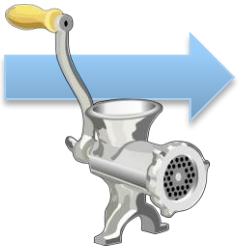
johnd@domainA:
{john, johnd,
joseph, doe}

Document-term matrix

	johnd@...	j.doe@...		
john	1	..	..	..
johnd	1	..	..	..
joseph	1	..	..	..
jdoe	?	..	..	..
doe	1	..	..	..

Algorithm of Kouters et al.

<John Doe, johnd@domainA>
<John Joseph Doe, johnd@domainA>



johnd@domainA:
{john, johnd,
joseph, doe}

Document-term matrix

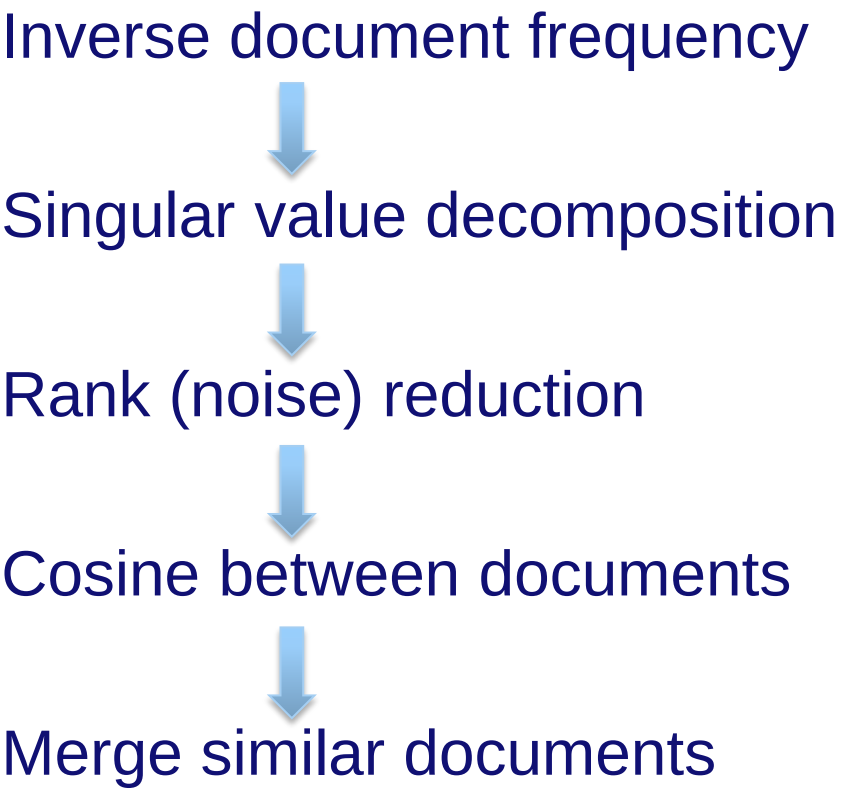
	johnd@...	j.doe@...		
john	1	..	..	..
johnd	1	..	..	..
joseph	1	..	..	..
jdoe	3/4	..	..	..
doe	1	..	..	..

max similarity(jdoe,
{john, johnd, joseph, doe})
= similarity(jdoe, doe)
= 1 – Levenshtein(jdoe, doe) /
max(length(jdoe), length(doe))
= 1 – 1/4 = 3/4

Latent Semantic Analysis

```
<John Smith, john@domainA>
<John Brown, john@domainB>
```

	johnd@...	j.doe@...		
john	1	..	..	..
johnd	1	..	..	..
joseph	1	..	..	..
jdoe	3/4	..	..	..
doe	1	..	..	..



Inverse document frequency

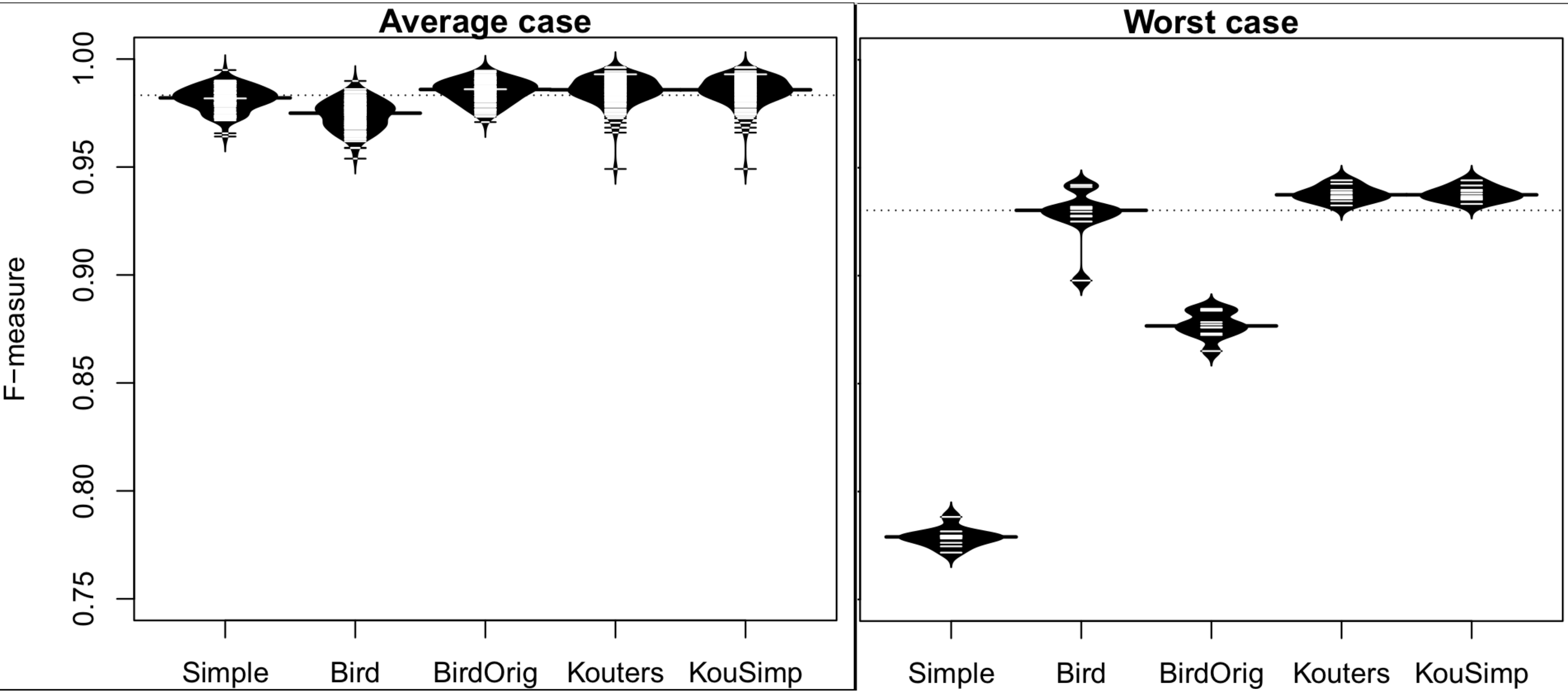
Singular value decomposition

Rank (noise) reduction

Cosine between documents

Merge similar documents

Empirical evaluation: GNOME



Identity merging: Summary

- Contributors use different aliases
  - In the same repository of across repositories

- Merging is needed for
  - Contributions, expertise, effort, social structure

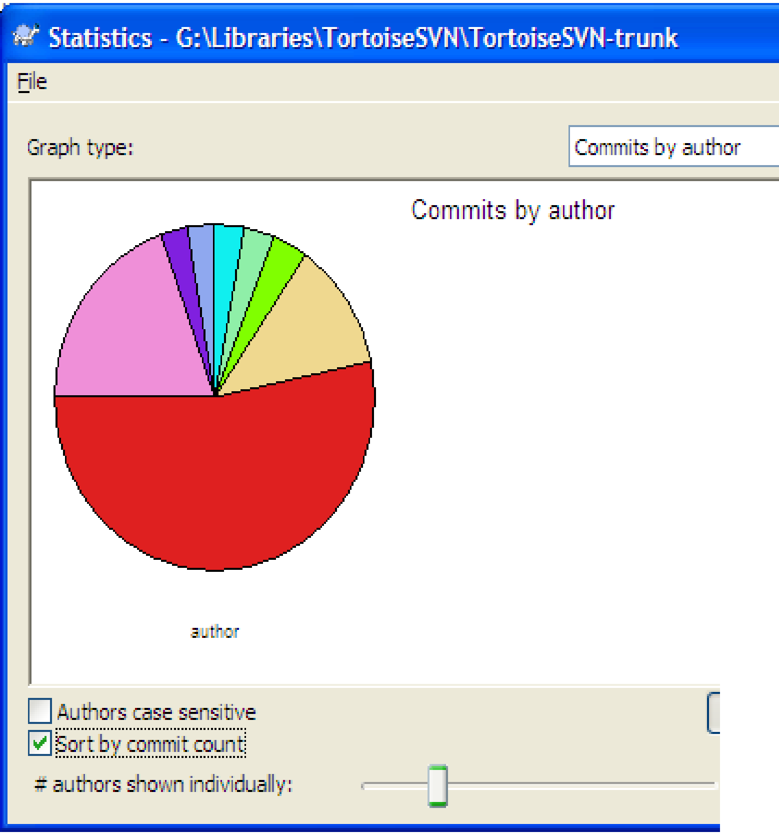
- Different merging algorithms
  - Simple, Bird’s, LSA

More research is needed…

- Different platforms ⇒ different kinds of noise ⇒ different techniques might be needed
- DBLP-like idea: people tend to work with the same partners on similar topics
- BUT… what about privacy?

What can we learn about the humans?

- Count commits per committer
- Look at how the counts evolve in time



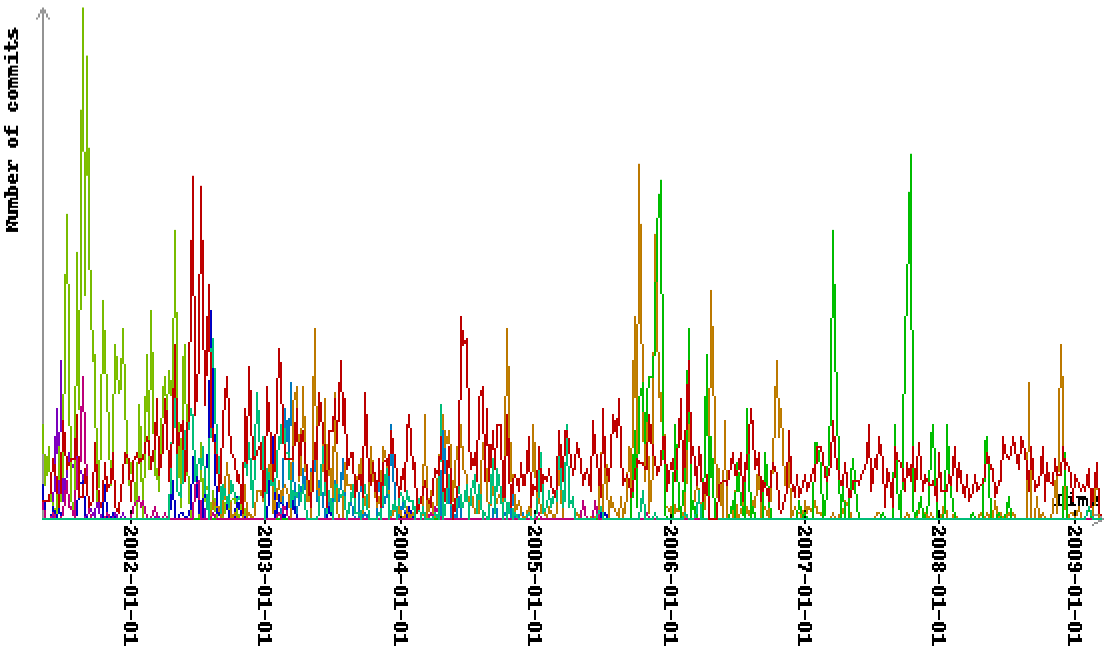
Statistics - G:\Libraries\TortoiseSVN\TortoiseSVN-trunk

Graph type: Commits by author

□ Authors case sensitive

☑ Sort by commit count

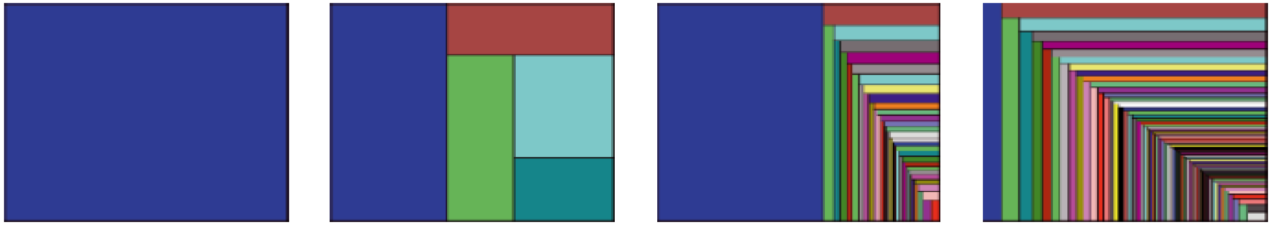
authors shown individually:



- One major committer?

More refined way of counting: Per File

- **What developer worked on a file**
  - **Count pc(Alice): the % of commits on F made by Alice**
  - **Visualization (Fractal Figure)**
    - **pc is a relative area of a rectangle**



(a) One developer

(b) Few balanced developers

(c) One major developer

(d) Many balanced developers

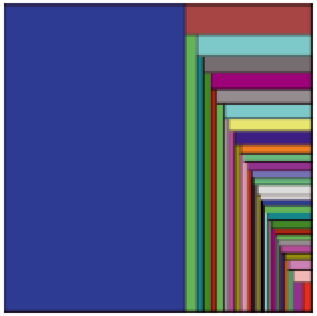
- **Measure of "difference"**

$$1-\sum_{c\in\text{committers}} pc^2(c)$$

- **How does this measure behave for (a), (b), (c) and (d)?**

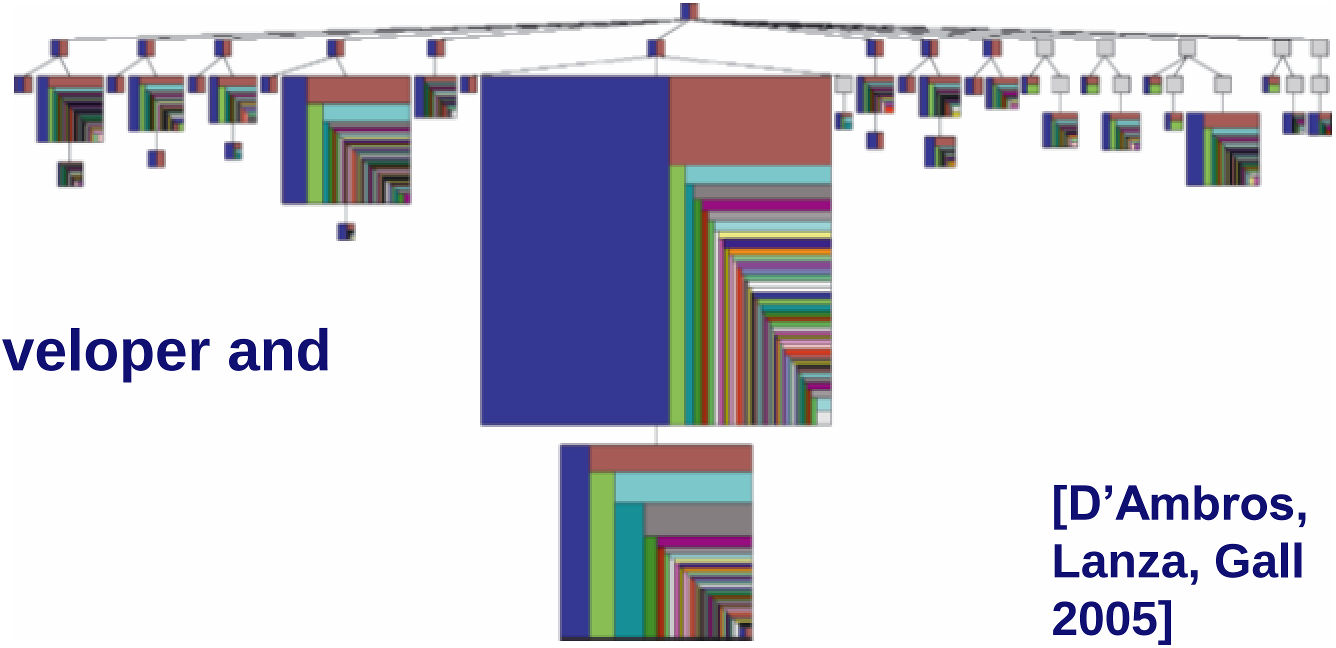
Fractal Figures

- pc is a relative area
  - Blue vs. red, green, …



- Many options for absolute size
  - Number of changes
  - Size of an artefact (file, directory)
  - Number of bugs

One major developer and many bugs!



[D’Ambros, Lanza, Gall 2005]

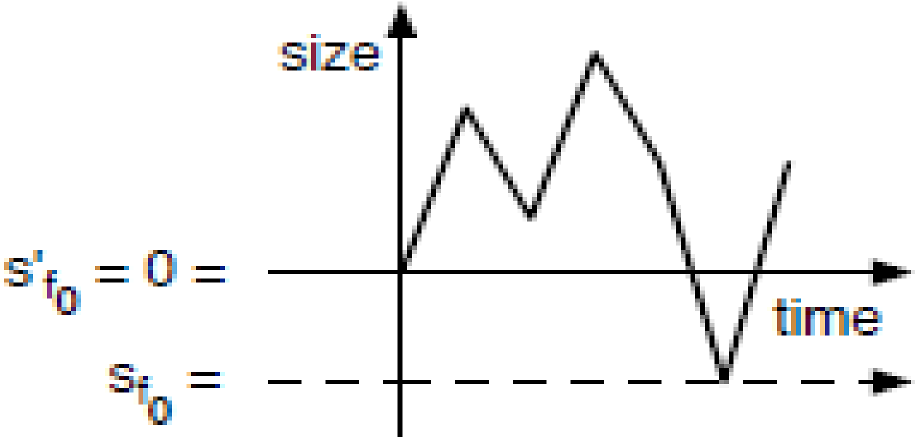
… Size of an artefact?

- **Easy to determine if the code is available**
- **Can be estimated if only the log is available [Gîrba Kuhn Seeberger Ducasse 05]**

≥ 8 lines before

≥30 lines after

```
Working file: insert-msg.tcl
…
revision 1.2
date: 1999/03/05 07:23:11; author: philg; state: Exp; lines: +30 -8
changed the bboard to do generic file uploading (and fixed Ben's
broken image uploading stuff)
```



$$s'_{f_0} := 0$$

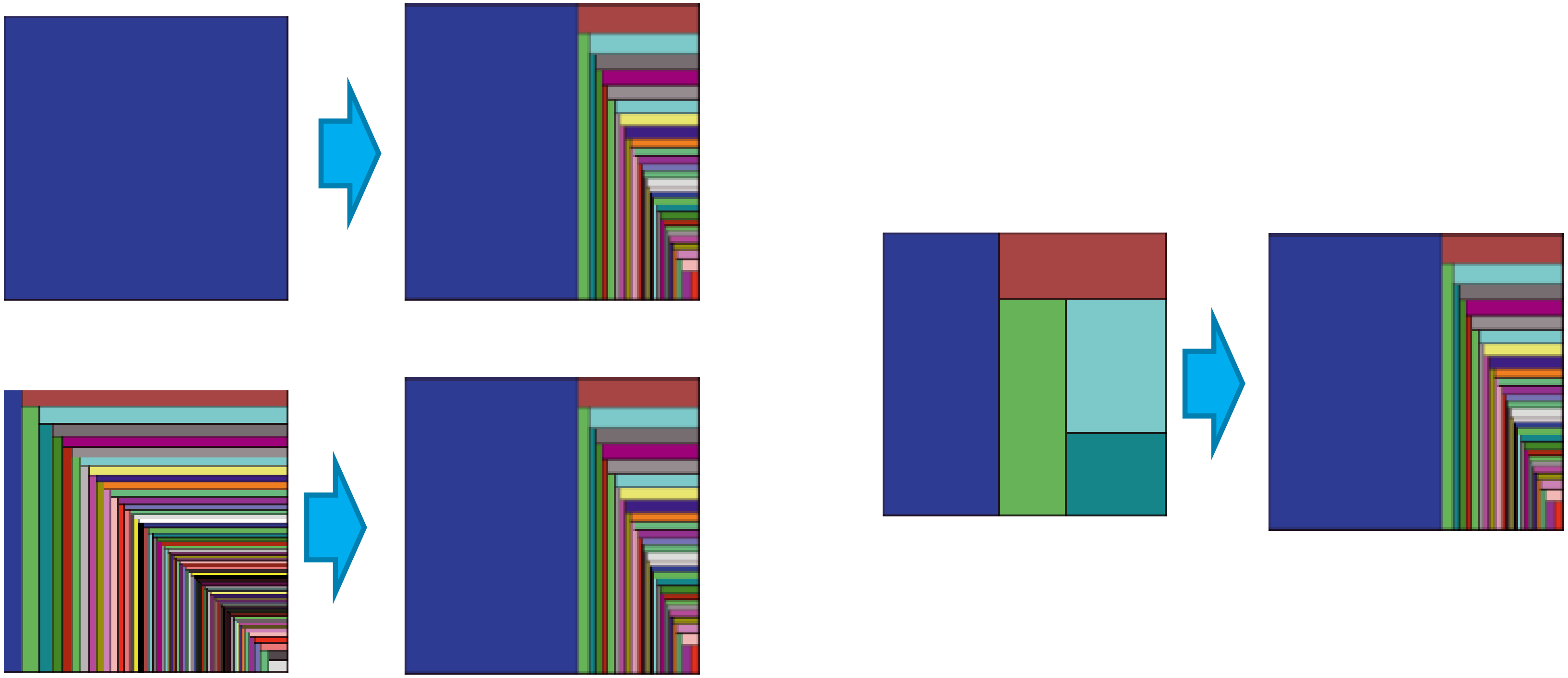
$$s'_{f_n} := s'_{f_{n-1}} + a_{f_{n-1}} - r_{f_n}$$

$$s_{f_0} := |min\{s'_x\}|$$

$$s_{f_n} := s_{f_{n-1}} + a_{f_n} - r_{f_n}$$

However we still have only a static view…

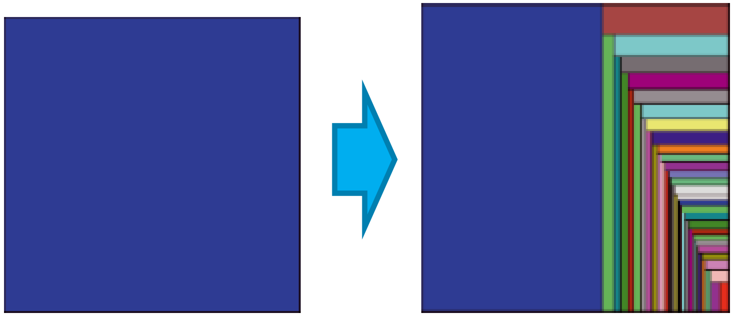
- How does the picture evolve in time?

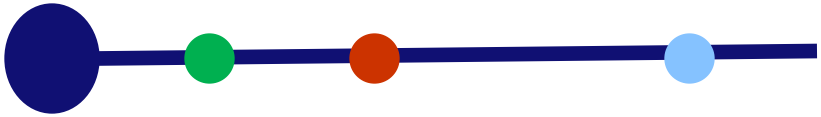


- Solutions:
  - Graph of fractal values
  - Ownership maps

Ownership maps [Gîrba Kuhn Seeberger Ducasse 05]

- Owner of…
  - line = last committer of this line
  - file = owns the major part of the lines
    - requires calculation of the file size
    - can be estimated from the log

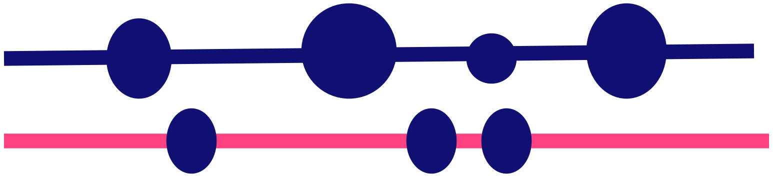




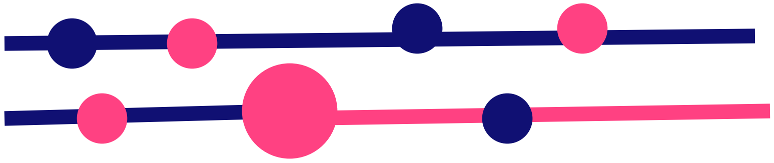
- Colour = committer
- Circle = commit
- Line = owner
- Timeline
- Size = proportion of change

Development patterns

- Monologue

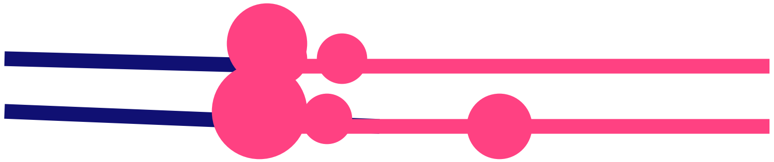


- Dialogue
  - Teamwork (quick succession)

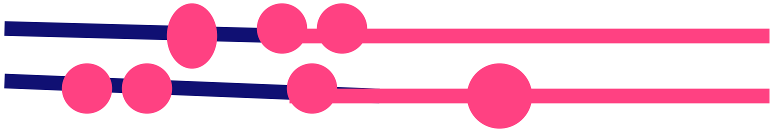


- Silence

- Takeover
  - Epilogue (Takeover + Silence)

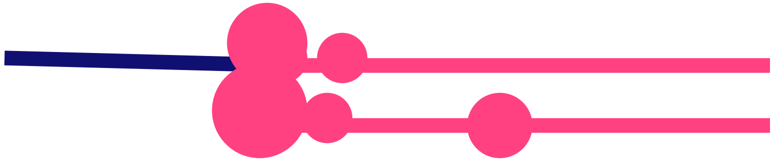


- Familiarization

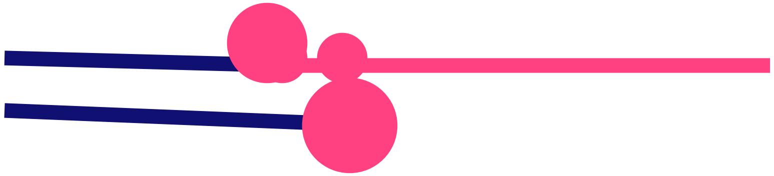


Development patterns (continued)

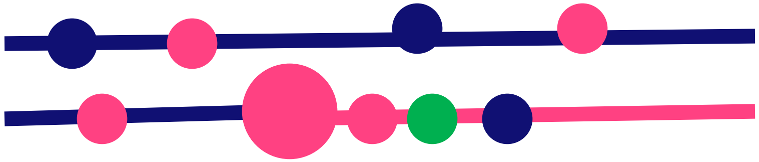
- **Expansion**



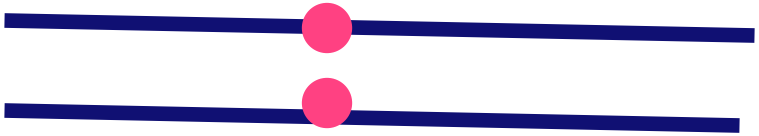
- **Cleaning**



- **Bug fix**



- **Edit**
  - **Epilogue (Edit + Silence)**

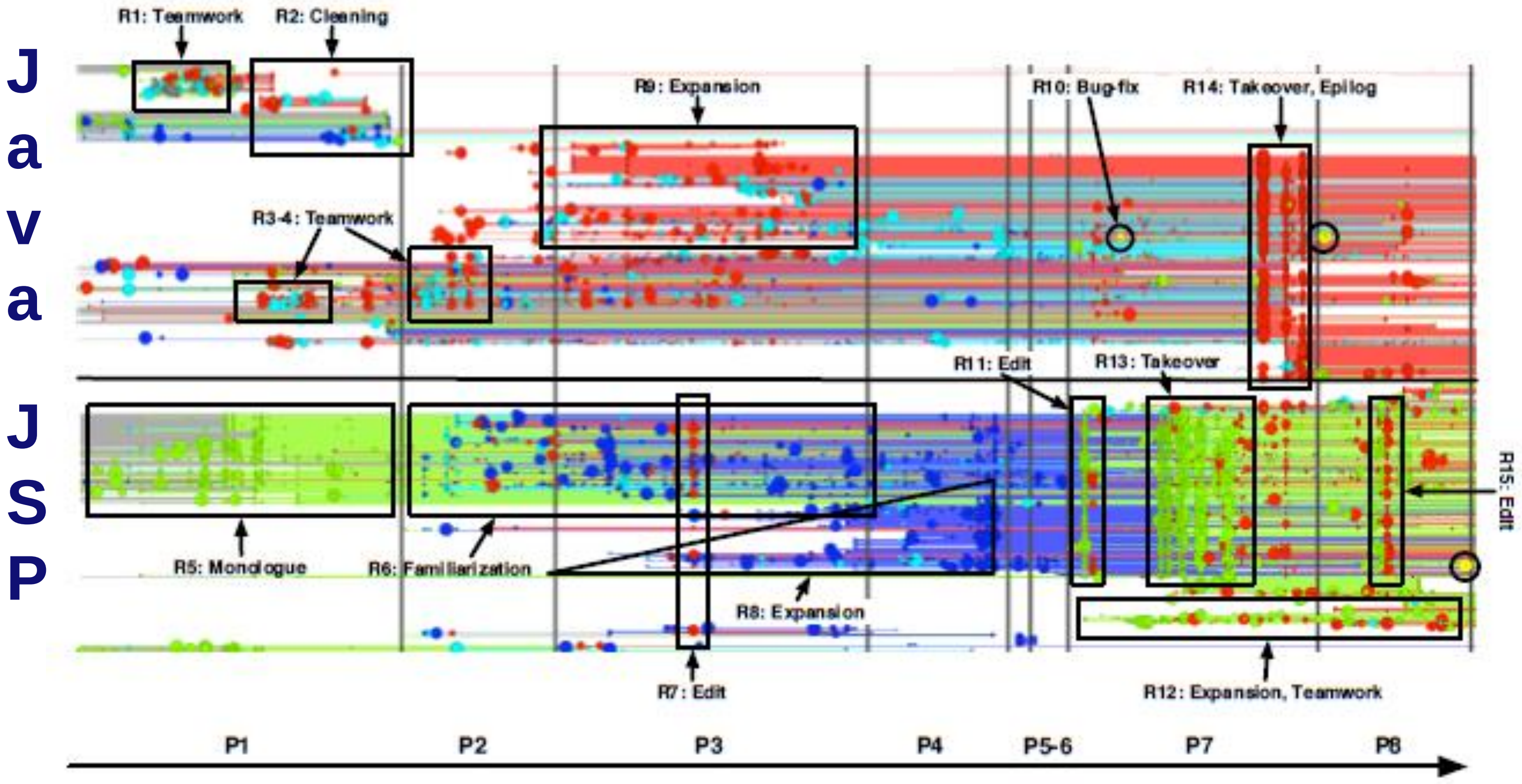


Experiment: Outsight

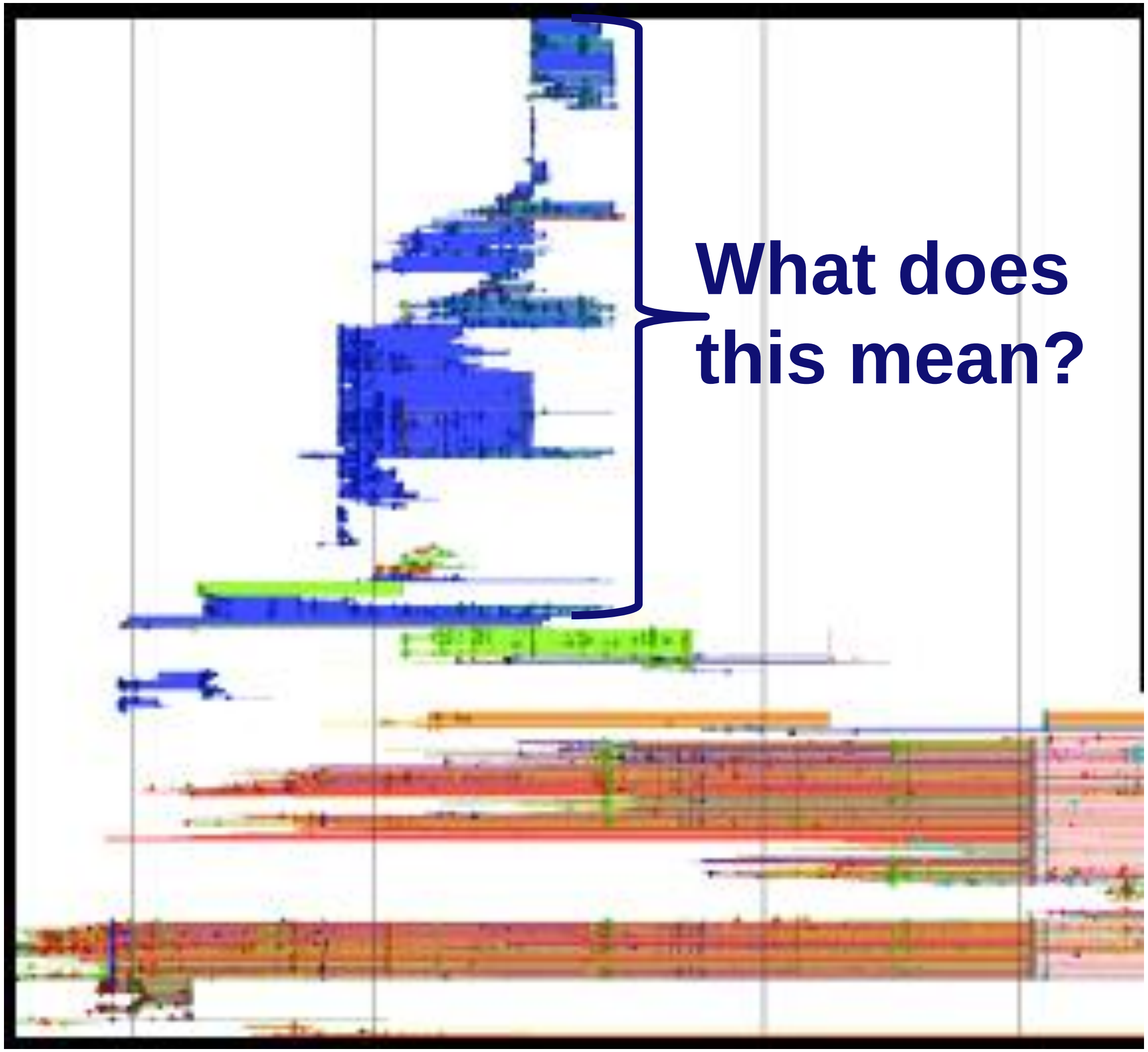
- Commercial application, 500 Java classes, 500 JSP
- 8 three-months periods

- How many developers are there?

- If you had questions about the system, whom would you ask?



Ant



What does this mean?

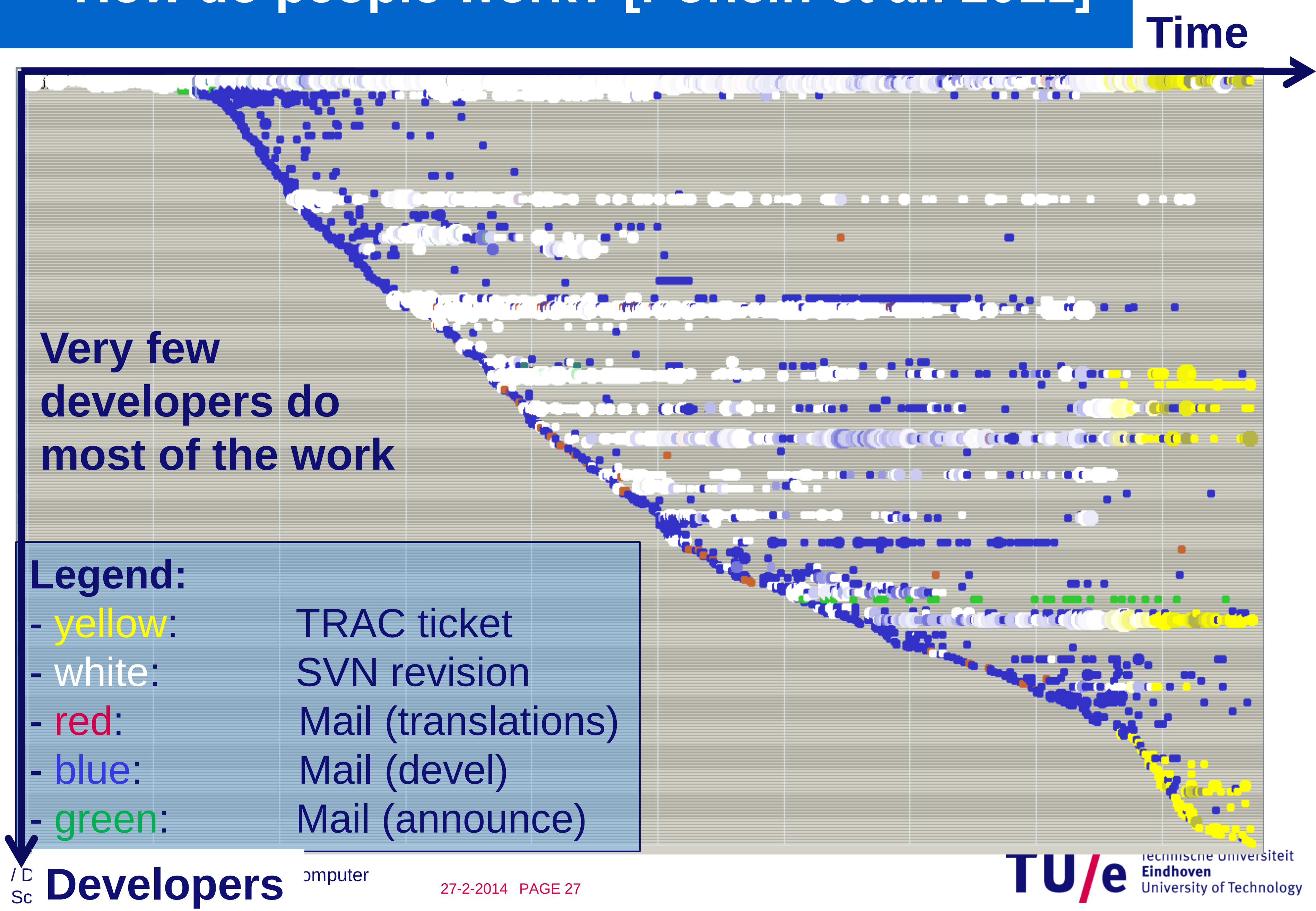
Subproject (Myrmidon) that was intended as a successor for Ant.

Pattern common to Open Source

Subprojects

- Cease
- Split
- Integrate in the main line

How do people work? [Poncin et al. 2011]



Very few developers do most of the work

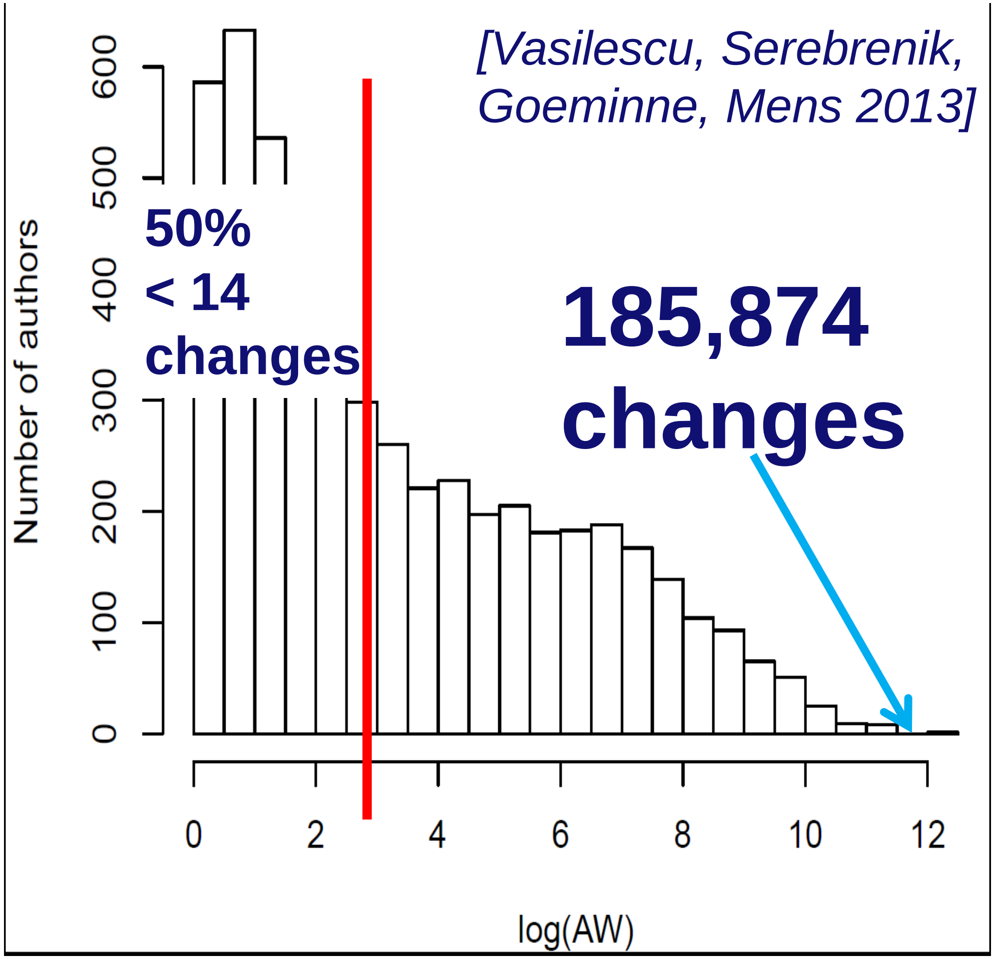
Legend:

- yellow:	TRAC ticket
- white:	SVN revision
- red:	Mail (translations)
- blue:	Mail (devel)
- green:	Mail (announce)

“Very few developers do most of the work”

- **GNOME**
  - **1316 projects**

- **NB: logarithmic scale on the x-axis**

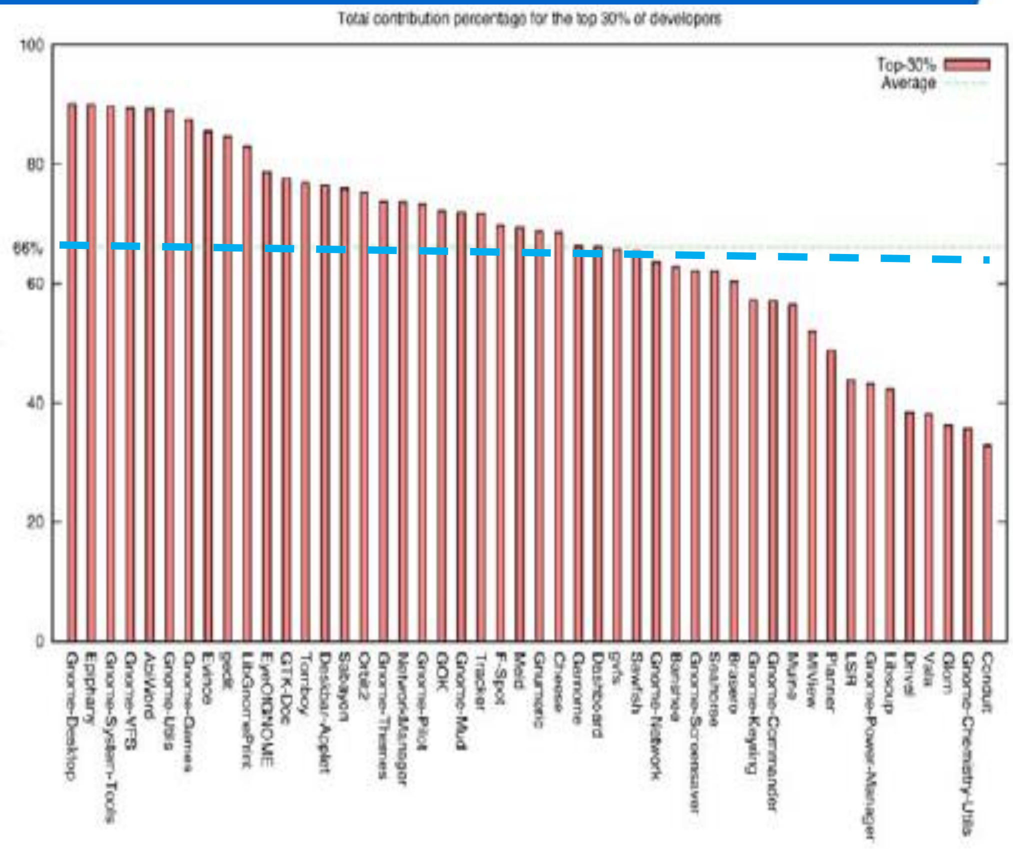


AW: number of changes of an author

“Very few developers do most of the work”

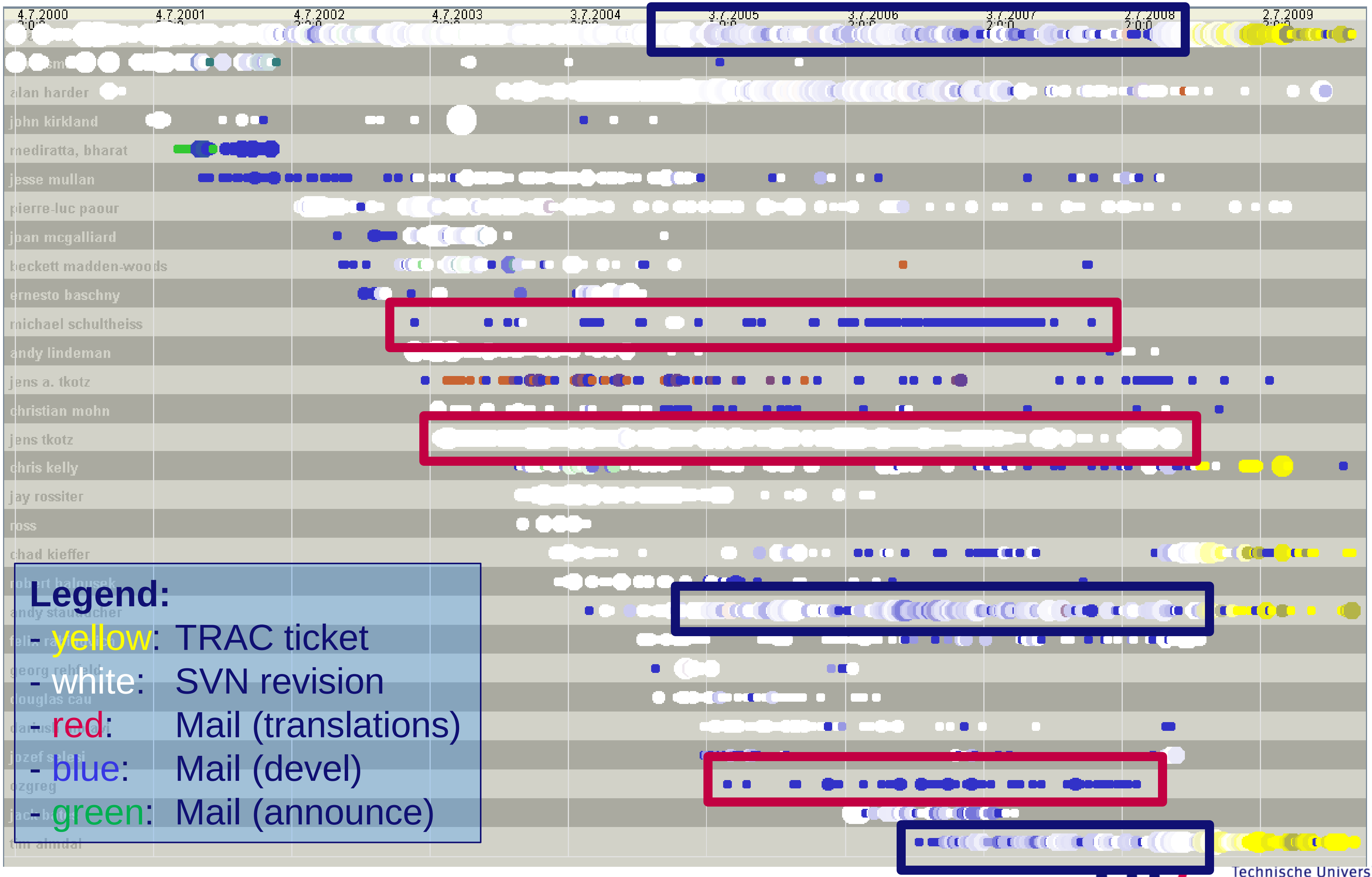
- **“Pareto principle” 20/80**

- **Quite common for software metrics**
  - **More precise descriptions of the distribution are possible**
  - **Even for LOC no agreement on the precise distribution**



Contribution of 30% most prolific developers in different GNOME projects [Kalliamvakou, Gousios, Spinellis, Pouloudi, 2009]

FRASR: Who does what?

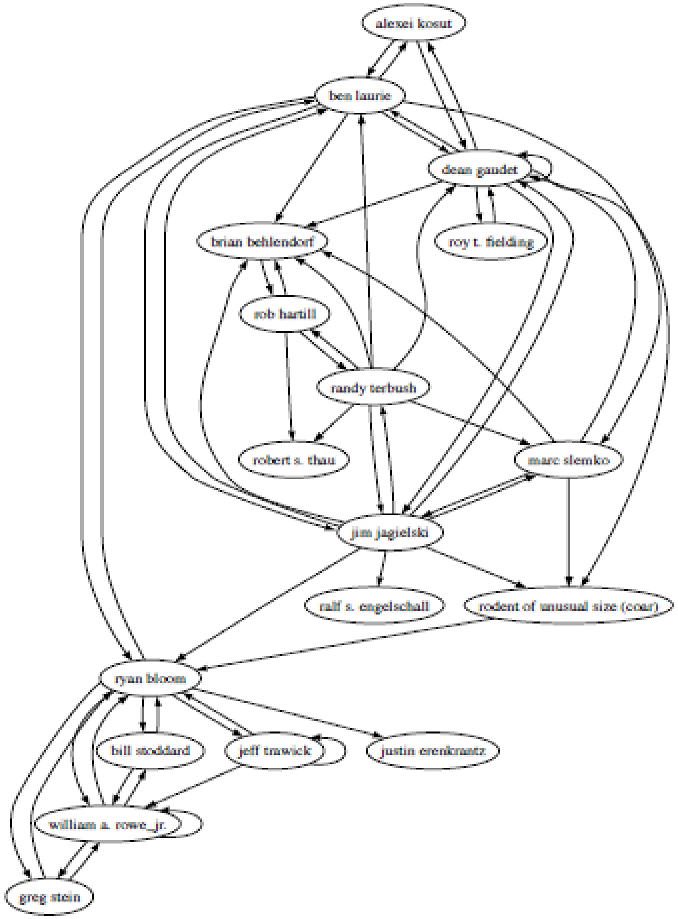


Legend:

- yellow: TRAC ticket
- white: SVN revision
- red: Mail (translations)
- blue: Mail (devel)
- green: Mail (announce)

All developers are equal, but some are more equal than others [Bird et al. 2006]

- Mail archive vs. version control
  - Without commit rights: “non-developers”
  - With commit rights: some commit more often

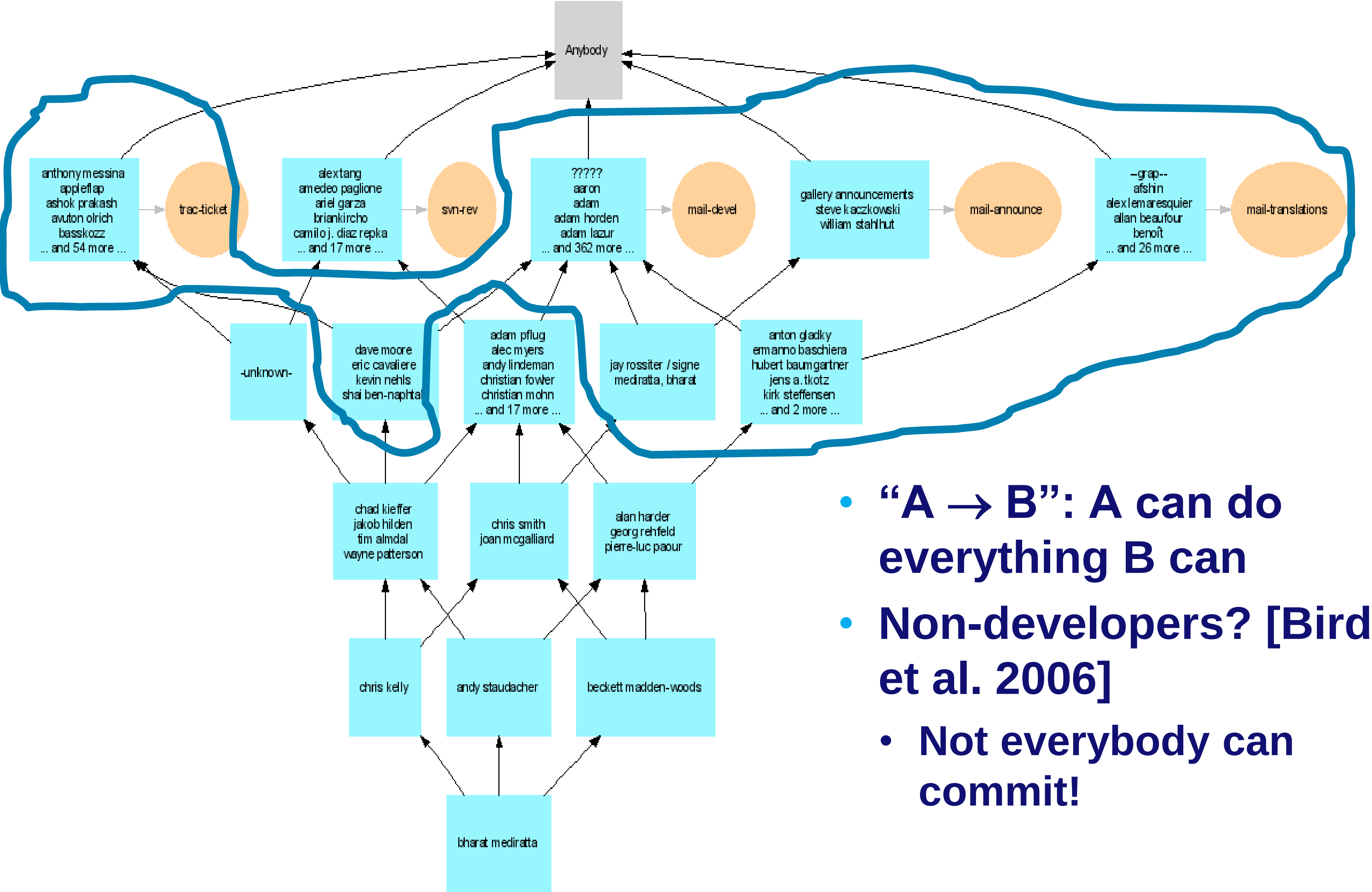


Mail communication (arrow = at least 150 mails send)

Conclusion 1: Developers are more active than non-developers

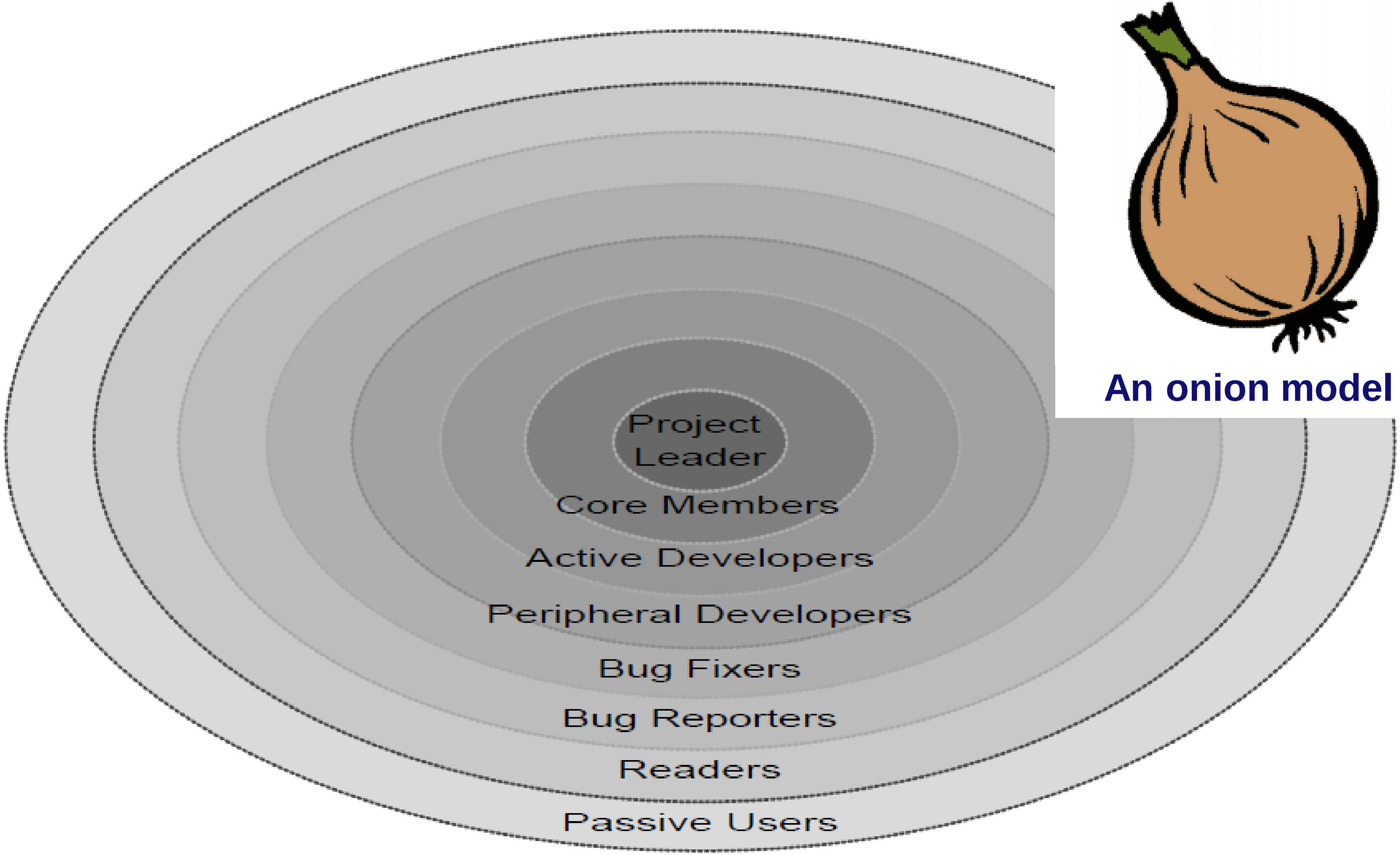
Conclusion 2: Correlation between the number of commits and the “centrality” of the developer

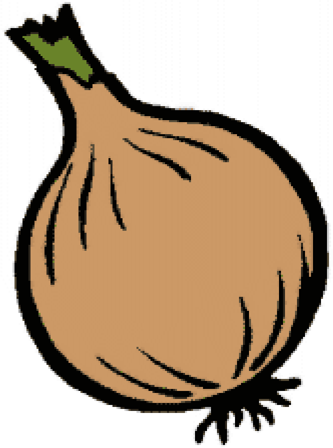
More refined developers classification is possible! [Wouter Poncin]



- **“A → B”: A can do everything B can**
- **Non-developers? [Bird et al. 2006]**
  - **Not everybody can commit!**

What kind of roles do the developers play?

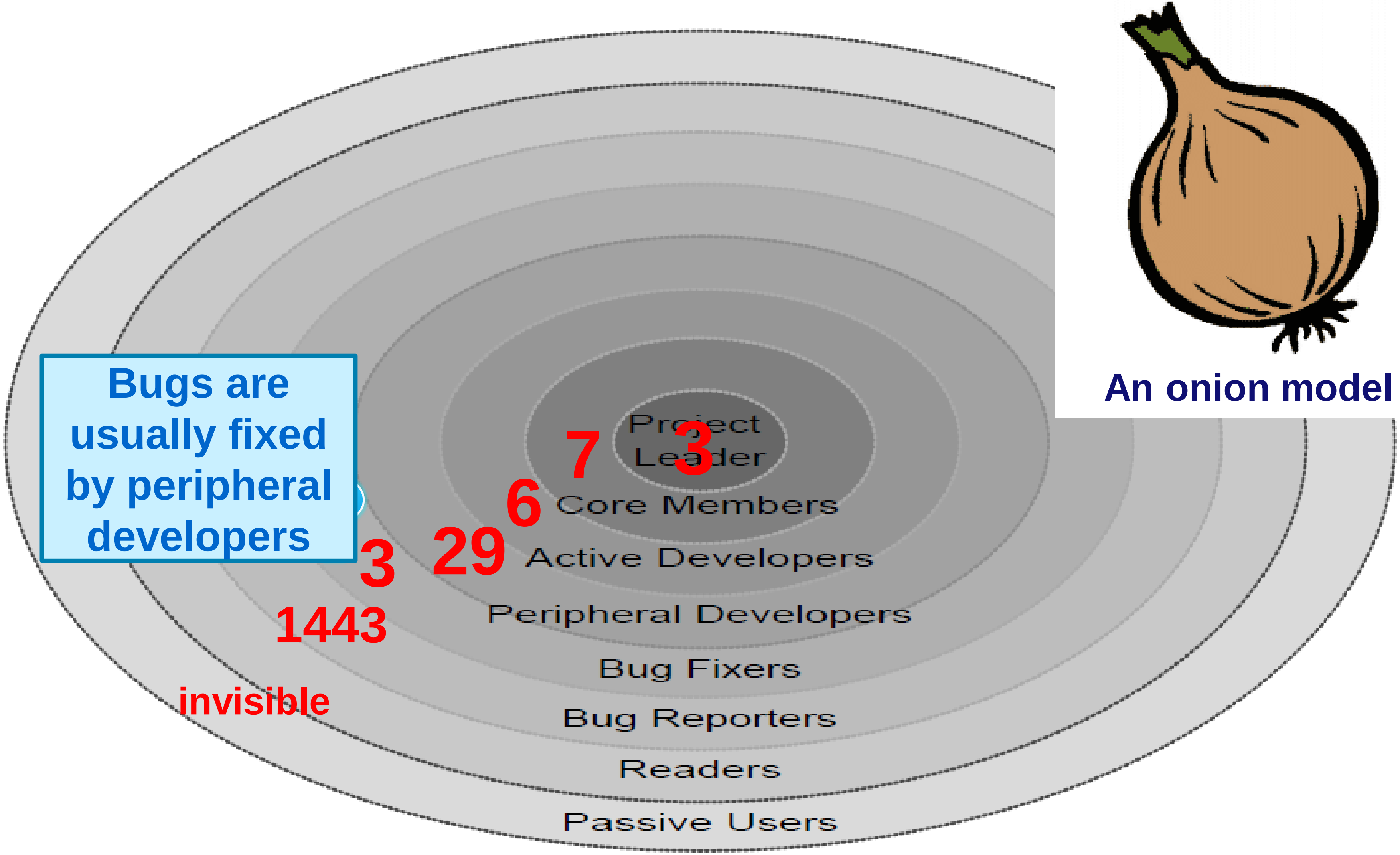




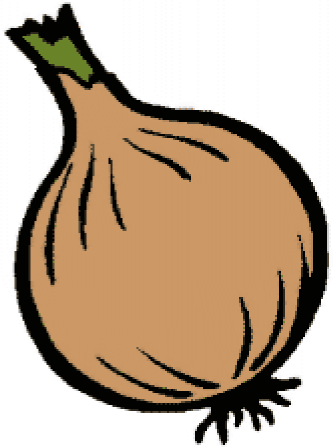
An onion model

Nakakoji et al. 2002

Onion in aMSN



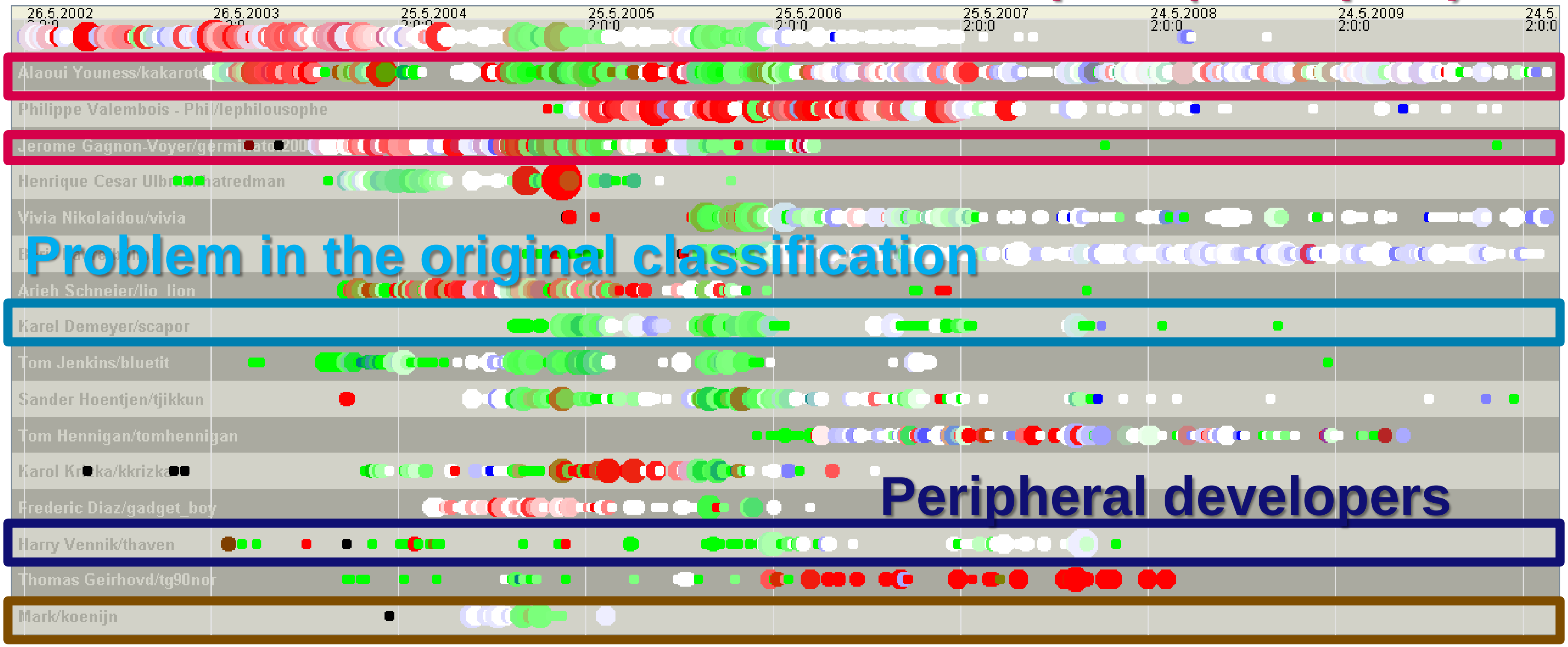
Bugs are usually fixed by peripheral developers



An onion model

Nakakoji et al. 2002

Nakakoji et al. as a case of

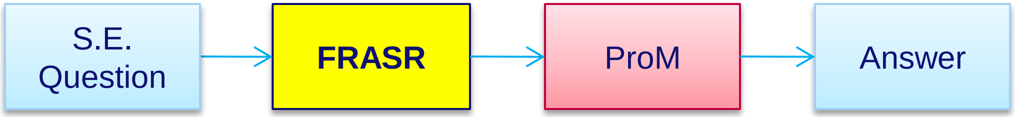


Core developers (examples)

Problem in the original classification

Peripheral developers

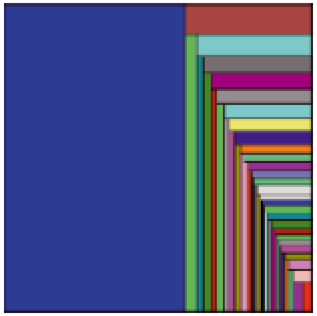
Bug reporter

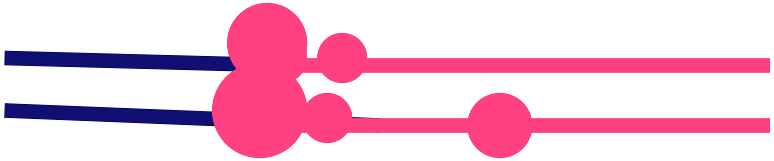


S.E. Question → FRASR → ProM → Answer

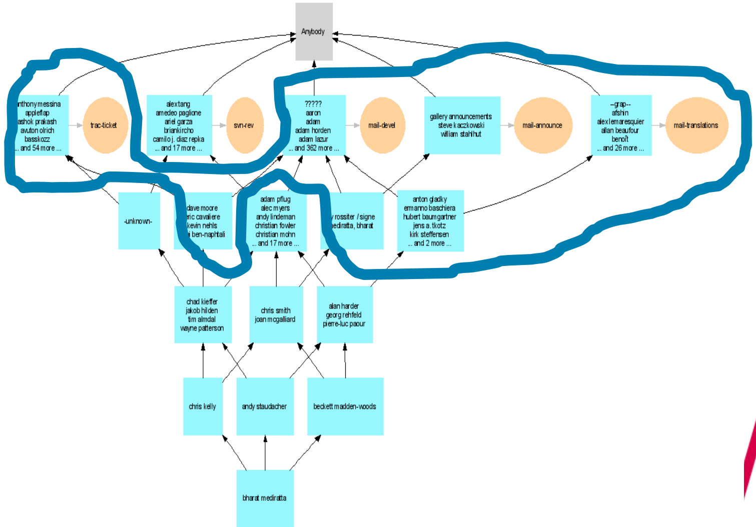
What can we learn about developers?

- Development effort distribution and evolution





- Can be combined with other information to distinguish different kinds of developers

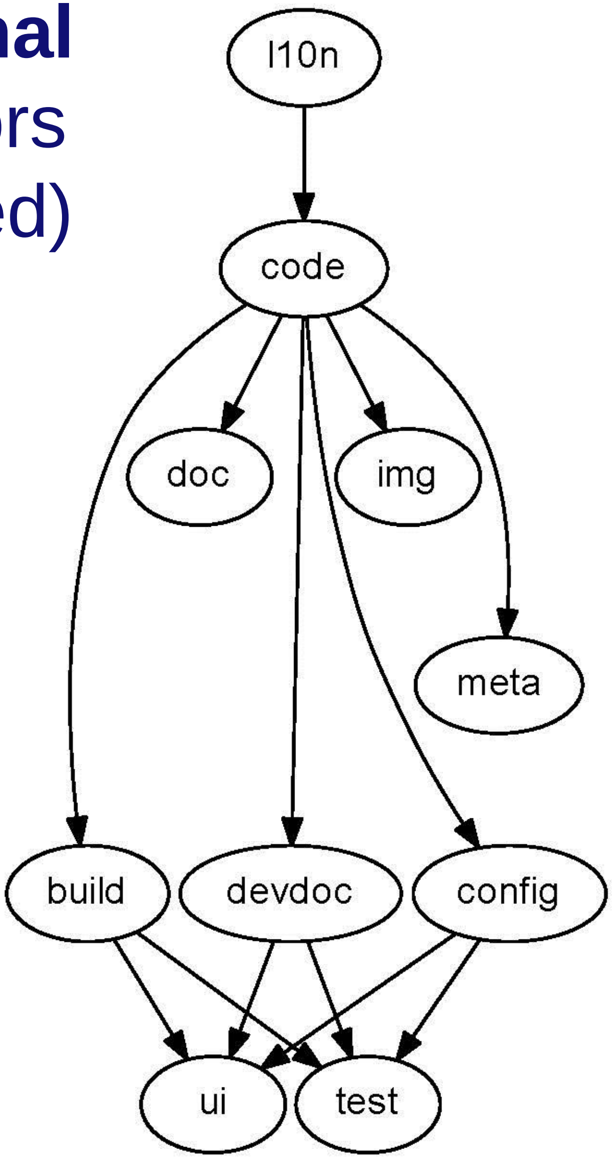


Not only developers

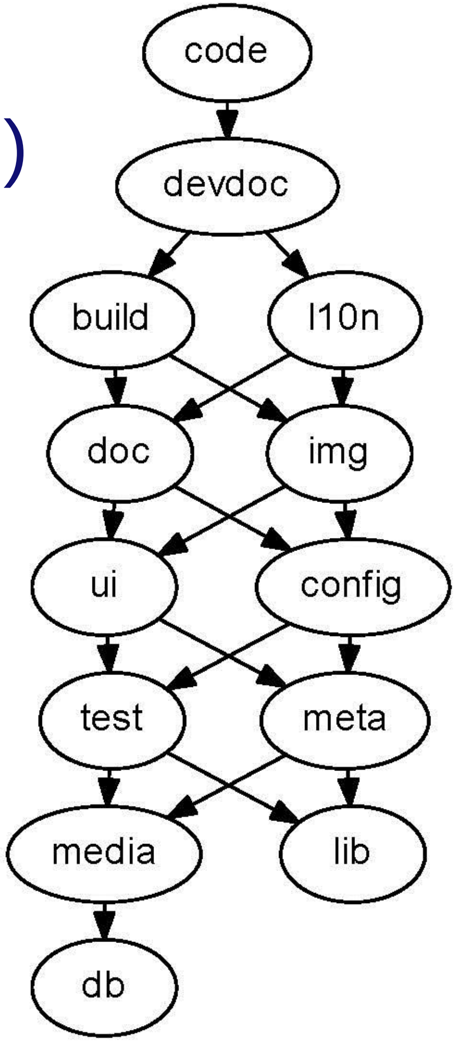
“Since 1997, the GNOME project has grown from a handful of developers to a contributor base of **coders**, **documenters**, **translators**, **interface designers**, **accessibility specialists**, **artists** and **testers** numbering in the thousands.“ (Waugh 2007)

Localization and coding

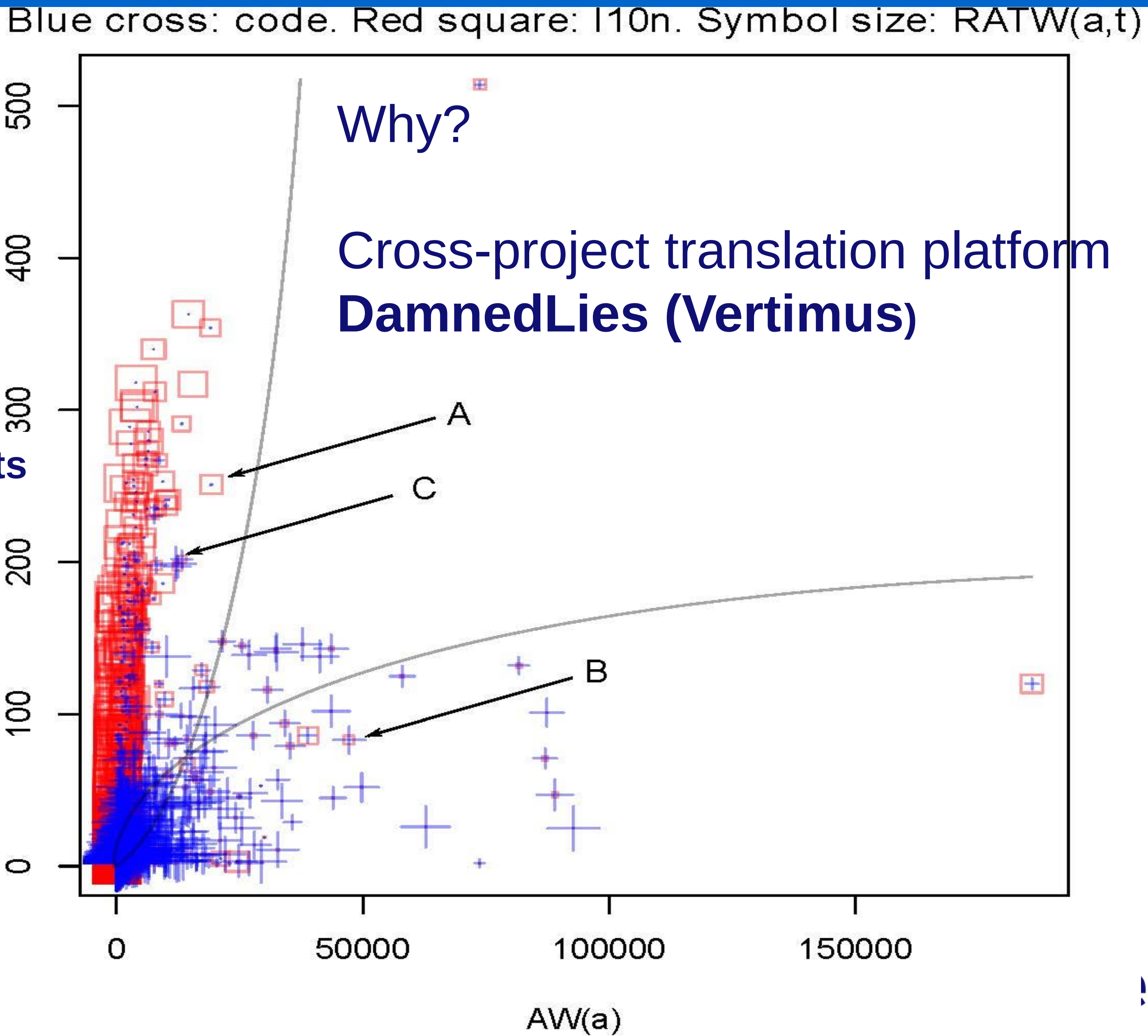
Occasional
contributors
(AW < med)



Frequent
contributors
(AW >= med)



Coding and localization in GNOME



Why?

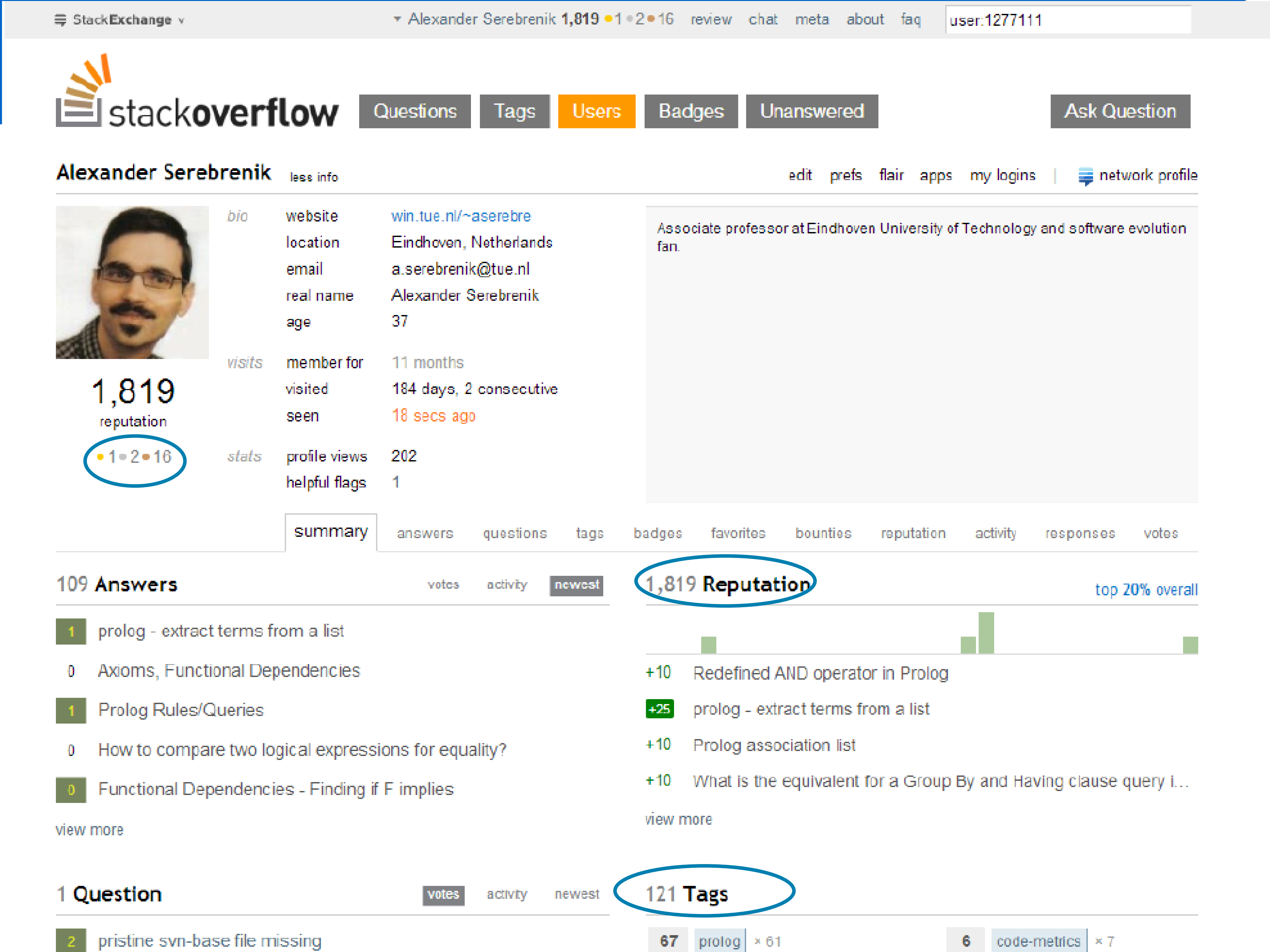
Cross-project translation platform **DamnedLies (Vertimus)**

#projects

Mythbusters

- **Once coder, always coder…**
  - **True** for coding and localization
  - **False** for, e.g., database development

- **Translation done in the target-language country is better!**
  - GNOME, French
  - In-country:
    - more translation mistakes
    - lower impact on understanding



StackExchange ▾ ▾ Alexander Serebrenik 1,819 ●1 ●2 ●16 review chat meta about faq user:1277111

stackoverflow Questions Tags Users Badges Unanswered Ask Question

Alexander Serebrenik less info

edit prefs flair apps my logins | network profile

bio	website	win.tue.nl/~aserebre
	location	Eindhoven, Netherlands
	email	a.serebrenik@tue.nl
	real name	Alexander Serebrenik
	age	37
visits	member for	11 months
	visited	184 days, 2 consecutive
	seen	18 secs ago
stats	profile views	202
	helpful flags	1

1,819
reputation
●1 ●2 ●16

Associate professor at Eindhoven University of Technology and software evolution fan.

summary answers questions tags badges favorites bounties reputation activity responses votes

109 Answers

votes activity newest

1	prolog - extract terms from a list
0	Axioms, Functional Dependencies
1	Prolog Rules/Queries
0	How to compare two logical expressions for equality?
0	Functional Dependencies - Finding if F implies

view more

1,819 Reputation

top 20% overall

+10	Redefined AND operator in Prolog
+25	prolog - extract terms from a list
+10	Prolog association list
+10	What is the equivalent for a Group By and Having clause query i...

view more

1 Question

votes activity newest

2	pristine svn-base file missing

121 Tags

67	prolog	× 61	6	code-metrics	× 7

Women and StackOverflow

http://meta.stackoverflow.com/questions/30411/

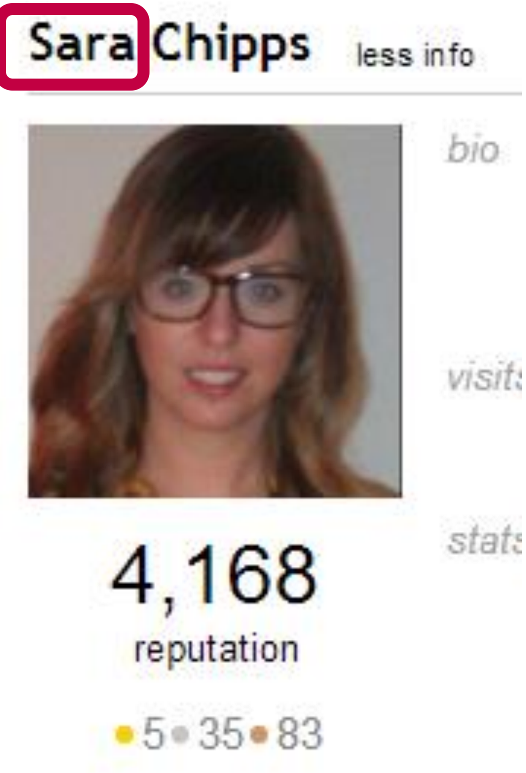
- **Ikessler**: I know a lot of female programmers, and I know there are a good number of them out there. But I don't recall ever having one of my questions answered by, nor have I ever answered a question by a female programmer here at Stack Overflow.

- **Sara Chipps**: there is NO appeal for me in answering questions.
- **Ether**: A huge number of SO users don't use their real names, so you actually have no idea.
- **Heather**:
  - Sexism still exists.
  - Women are still perceived as lightweights.

Women, men, StackOverflow and more

- Our questions:
  - Did women really participate in SO less than men?
    - random sample
  - Is this SO specific?
    - Compare with Drupal and Wordpress mailing lists
- But first: *what is your gender?*

What is your gender?



Sara Chipps less info

bio

visits

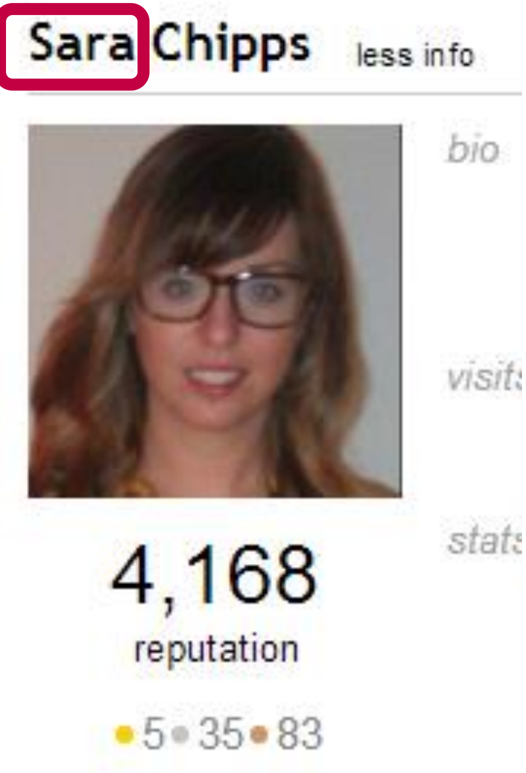
stats

4,168

reputation

● 5 ● 35 ● 83

What is your gender?



Sara Chipps less info

bio

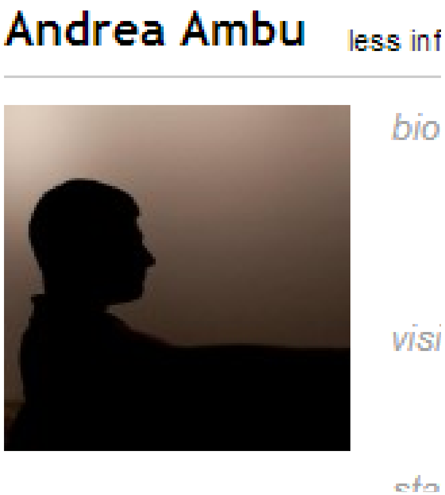
visits

stats

4,168

reputation

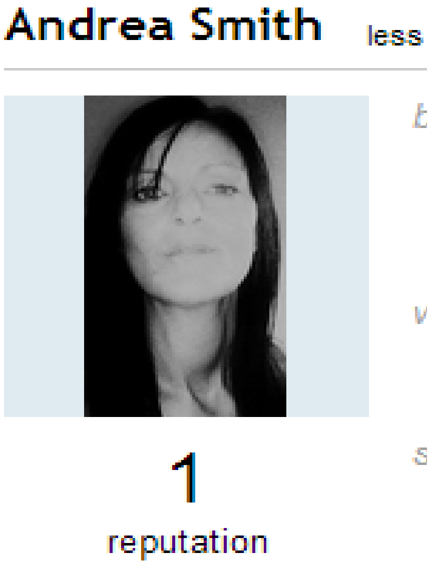
●5 ●35 ●83



Andrea Ambu less inf

bio

visi

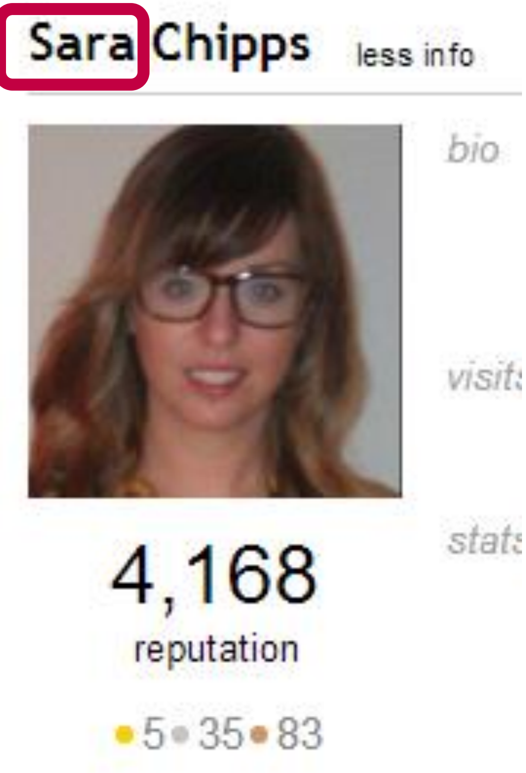


Andrea Smith less

1

reputation

What is your gender?



Sara Chipps less info

bio

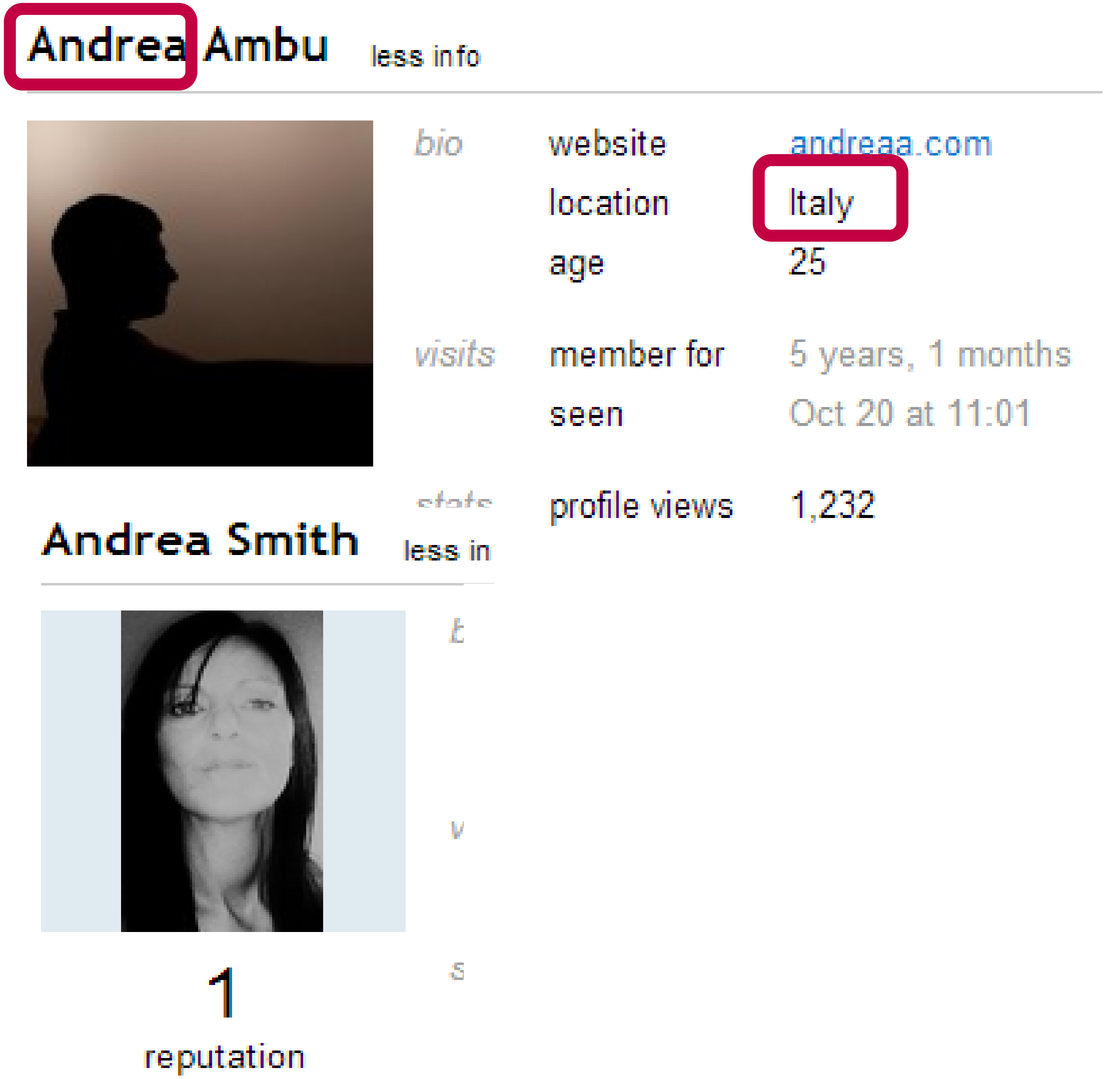
visit

stat

4,168

reputation

●5●35●83



Andrea Ambu less info

bio	website	andreaa.com
	location	Italy
	age	25
visits	member for	5 years, 1 months
	seen	Oct 20 at 11:01
stats	profile views	1,232

Andrea Smith less in

b

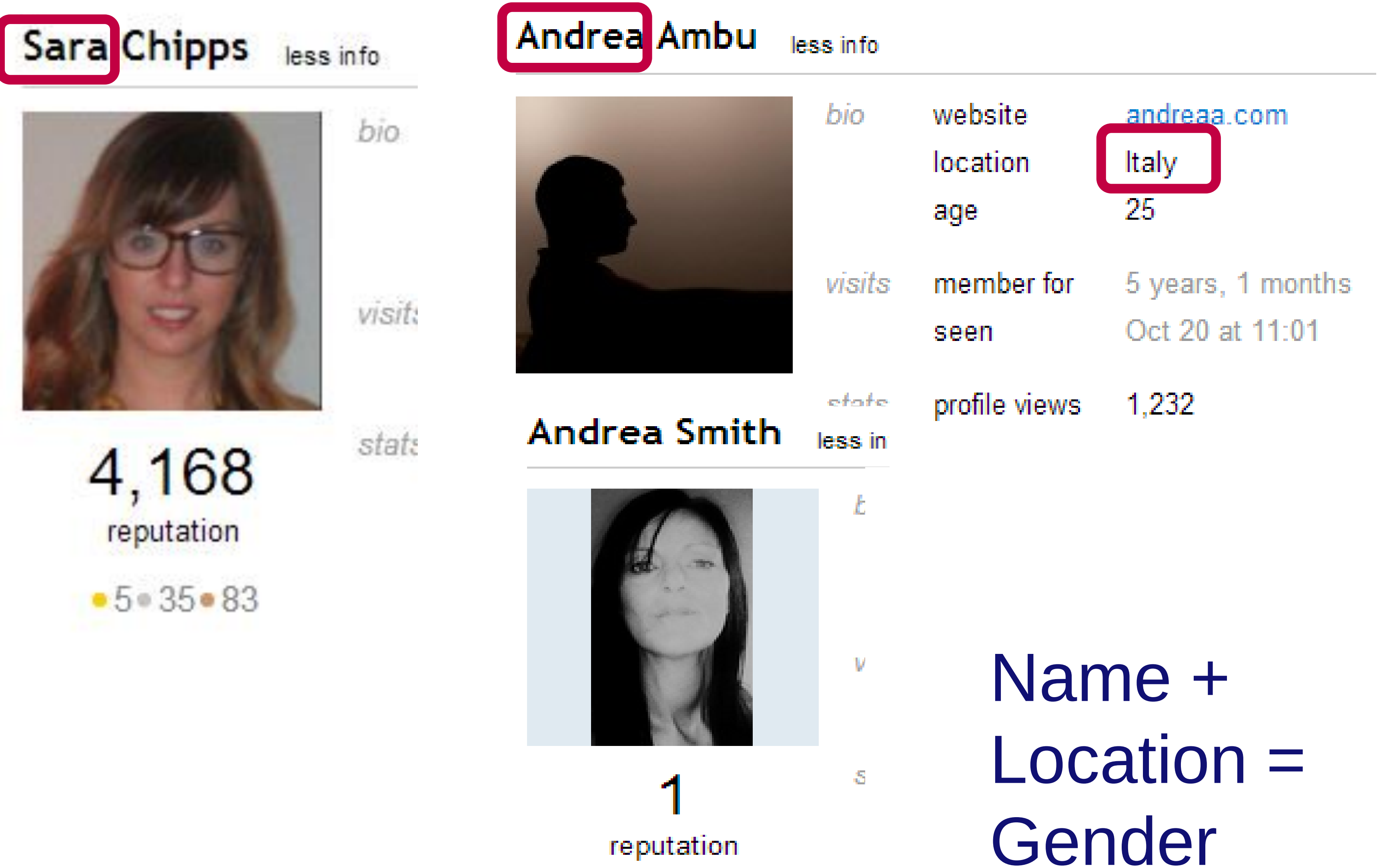
v

s

1

reputation

What is your gender?



Sara Chipps less info

bio

visits

stats

4,168
reputation

●5 ●35 ●83

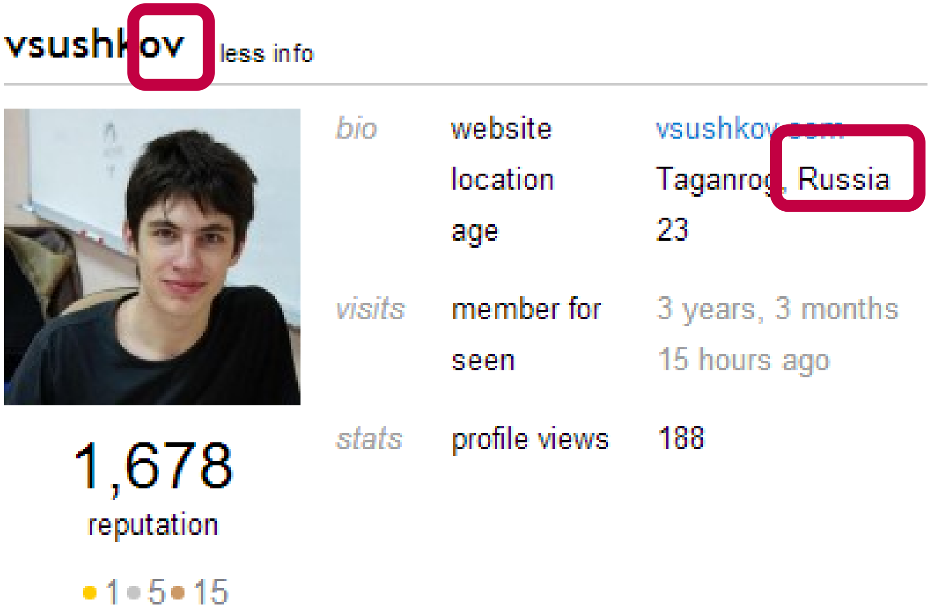
Andrea Ambu less info

bio	website	andreaa.com
	location	Italy
	age	25
visits	member for	5 years, 1 months
	seen	Oct 20 at 11:01
stats	profile views	1,232

Andrea Smith less in

1
reputation

Name + Location = Gender

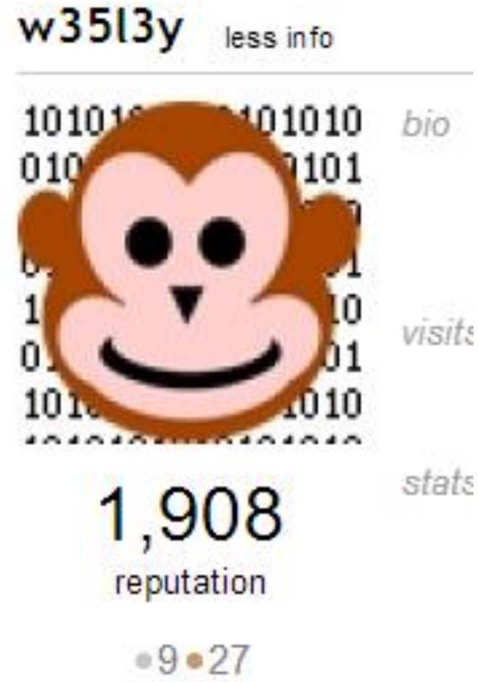


vsushkov less info

bio	website	vsushkov.com
	location	Taganrog, Russia
	age	23
visits	member for	3 years, 3 months
	seen	15 hours ago
stats	profile views	188

1,678
reputation

● 1 ● 5 ● 15



w35l3y less info

bio

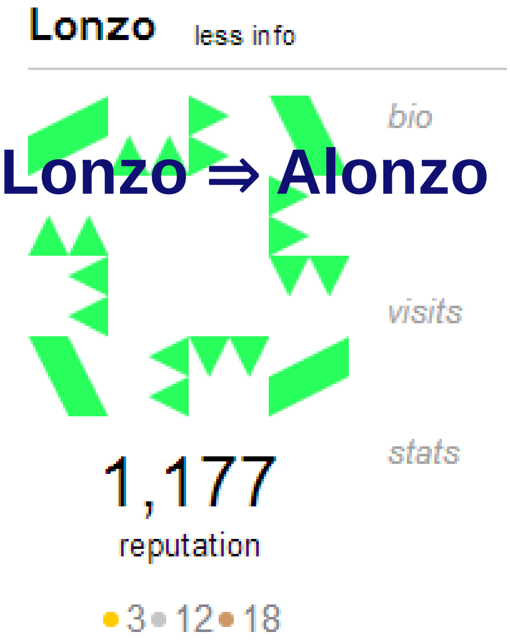
visits

stats

1,908
reputation

● 9 ● 27

w35l3y ⇒ wesley



Lonzo less info

bio

visits

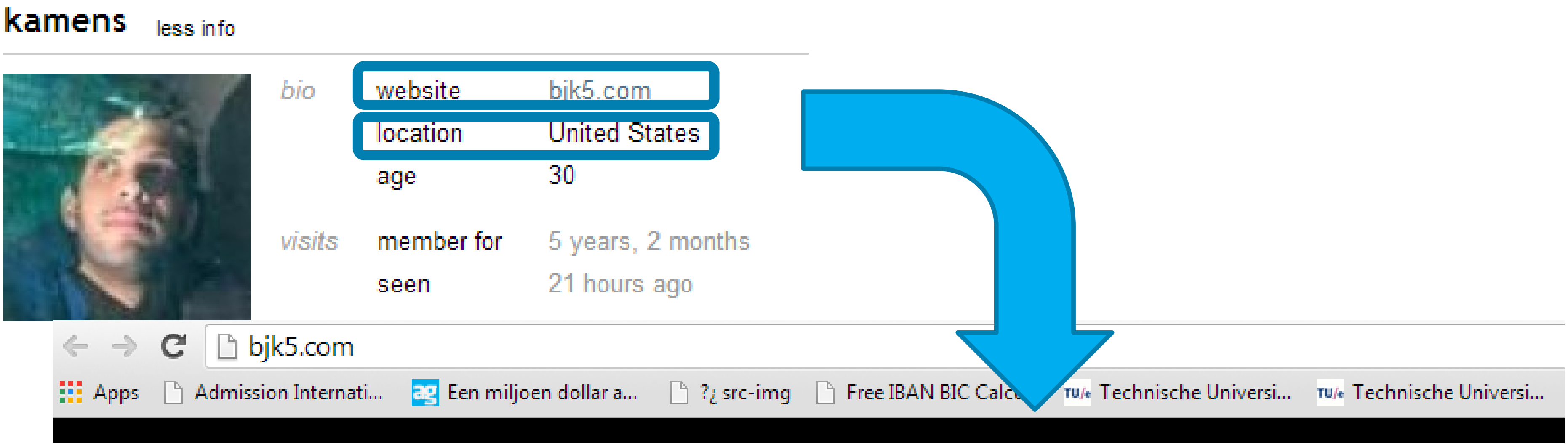
stats

1,177
reputation

● 3 ● 12 ● 18

Lonzo ⇒ Alonzo

Name + Location = Gender

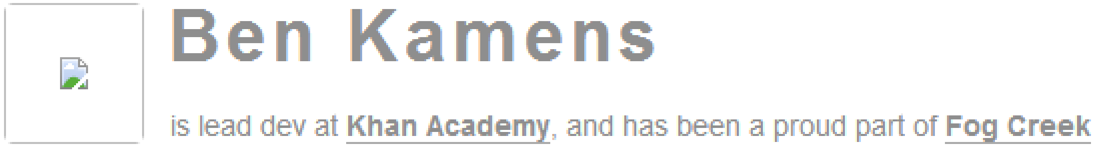


kamens less info

bio	website	bik5.com
	location	United States
	age	30
visits	member for	5 years, 2 months
	seen	21 hours ago

bjk5.com

Heuristics:
title + first h1



Ben Kamens

is lead dev at **Khan Academy**, and has been a proud part of **Fog Creek**

<title>Ben Kamens</title>
…
<h1>We&#8217;re willing to be embarrassed about what we <em>haven&#8217;t</em> done&#8230;</h1>

Ben Kamens We’re willing to be embarrassed about what we haven’t done…

Stanford Named Entity Tagger

<PERSON>Ben Kamens</PERSON> We’re willing to be embarrassed about what we haven’t done…

Quality of gender resolution: Survey

Self-identification	As inferred			Total
	M	F	?	
M	**60**	3	*43*	106
F	2	**5**	*4*	11

+ avatars, other social media sites (manually)

Self-identification	As inferred			Total
	M	F	?	
M	**90**	3	*13*	106
F	2	**9**	*0*	11

stack overflow *sample*	**2296**	**291**	**1557**
WordPress	**3043**	**282**	**286**
Drupal	**2879**	**328**	**135**

stack overflow *sample*	2296	291	1557
WordPress	3043	282	286
Drupal	2879	328	135

7-10% women as opposed to
1-5% for Open Source and
up to 28% for proprietary

stack overflow *sample*	**2296**	**291**	**1557**
WordPress	**3043**	**282**	**286**
Drupal	**2879**	**328**	**135**

7-10% on **different** mailing lists
more on “use technology”
less on “design technology”

stack overflow *sample*	**2296**	**291**	**1557**
WordPress	**3043**	**282**	**286**
Drupal	**2879**	**328**	**135**

It is easy to remain anonymous on SO and participants use this opportunity (**37.5%**)

stack overflow

sample

WordPress

Drupal

No significant differences in #questions, #answers, length of engagement

Affects eng’t for “design tech.” lists

stack overflow

sample

WordPress

Drupal

Engage for longer

Ask more questions

No diff in #answers

Women can contribute to SO but choose not to!

Why?

- [Gneezy, Niederle, Rustichini 2003]: women are less effective in mixed-gender competitive environments

- [Niederle, Vesterlund 2007]: women shy away from competition and men embrace it

⇒ To retain women we need **different gamification techniques**

Sounds interesting? Talk to me!

Capita Selecta opportunities

Conclusions

- Software repositories
  - Mail archives, version control, StackOverflow…

- Technical challenge: identity merging

- We can discover information about:
  - Roles (a la Nakakoji)
  - Activities (localization, coding, …)
  - Gender
  - Communication patterns
  - But also: age, location, culture, psychological type…