Risky Files: An Approach to Focus Quality Improvement Effort

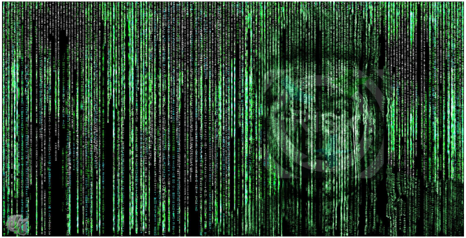
Audris Mockus Randy Hackbarth John Palframan

Avaya Labs Research
211 Mt Airy Rd
Basking Ridge, NJ 07920
audris@avaya.com

Aug 21, 2013

Motivation

Make *quality* of the code *transparent*



Indications

- Development transferred
- Few original authors remain
- A long development history
- Many customers/customer issues
- A component of many projects

Motivation

Make *quality* of the code *transparent*

Indications

- Development transferred
- Few original authors remain
- A long development history
- Many customers/customer issues
- A component of many projects

Motivation

Make *quality* of the code *transparent*

Indications

- Development transferred
- Few original authors remain
- A long development history
- Many customers/customer issues
- A component of many projects

Benefits

Top 1% of all files contribute to 60+% of field defects

Make Transparent

- Where to rebuild lost expertise
- Where to focus quality improvement

Provide guidance for

- Cost effective actions
- Practices to reduce future defects

Approach Outline

- Data processing
  - Accessing data sources
  - Linking data sources
  - Obtaining risk predictors
- Prioritized candidate list
  - Details needed for action
    - Related files
    - Modification Requests (MRs)
    - Customer Reported Defects (CFDs)
    - Developer expertise
  - Determine and schedule actions
- Monitor actions and measure quality improvement

Data Sources

- Code changes
  - 1K+ projects using git/svn/ClearCase/SCCS/other VCS
  - 50M+ changes

Data Sources

- Code changes
  - 1K+ projects using git/svn/ClearCase/SCCS/other VCS
  - 50M+ changes
- MRs: Why change was made?
  - ClearQuest/JIRA/other: 1.6M MRs

Data Sources

- Code changes
  - 1K+ projects using git/svn/ClearCase/SCCS/other VCS
  - 50M+ changes
- MRs: Why change was made?
  - ClearQuest/JIRA/other: 1.6M MRs
- Support: which MRs came from users (CFDs)?
  - Customer support (Siebel)
- Directory: who represents that login?
  - Corporate directory
  - Yellow pages to map login to corporate handle

Linking Data

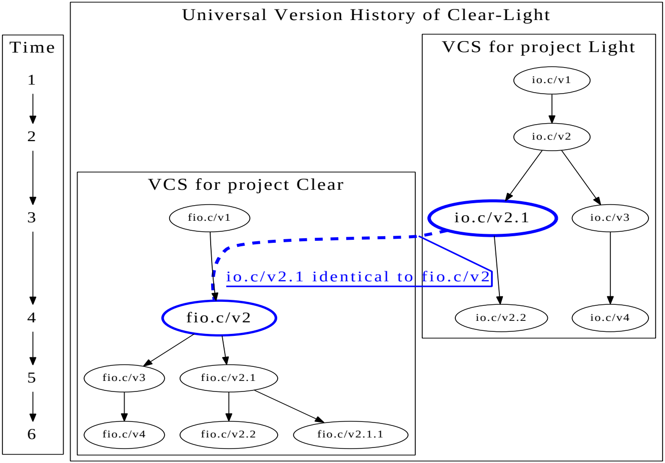
- MRs from code commit comments
- Corporate handle for commit login
- CFDs from Siebel

Linking Data

- MRs from code commit comments
- Corporate handle for commit login
- CFDs from Siebel

- Identify related (copied in the past) files
  - f_1 is directly related ($\sim$) to f_2 if
    $\exists v_1, v_2 : f_1(v_1) = f_2(v_2)$
    where $f(v)$ is a string representing version v of file f
  - f_1 is related to f_2 (a transitive closure of $\sim$) iff
    $\exists F_1, \ldots, F_k : f_1 \sim F_1, F_1 \sim F_2, \ldots, F_k \sim f_2$

io.c ∼ fio.c: directly related files



Determine risk factors most strongly associated with future customer-reported defects

Identify and prioritize files (equivalence classes)

- Risk predictors
  - Number of changes, CFDs
  - Number of authors, number who left
  - Size in LOC
  - Author experience
  - Number of static analysis warnings
  - % test coverage
- Risk prioritization
  - Fit a logistic regression model
  - Use a minimal subset to prioritize
- Produce top 1% risky file report

Determine risk factors most strongly associated with future customer-reported defects

Identify and prioritize files (equivalence classes)

- Risk predictors
  - **Number of changes, CFDs**
  - Number of authors, **number who left**
  - Size in LOC
  - Author experience
  - Number of static analysis warnings
  - % test coverage
- Risk prioritization
  - Fit a logistic regression model
  - Use a **minimal subset** to prioritize
- Produce top 1% risky file report

For subject matter experts (SMEs)

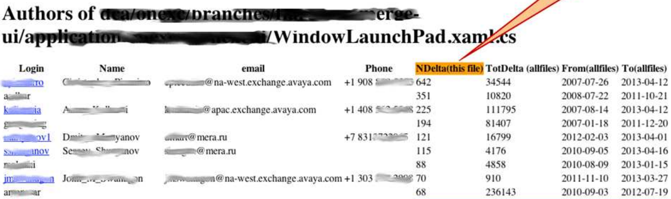
- In three formats
  - Hypertext, sortable by metrics, CSV
- Hypertext: for each file
  - Link to related files
  - Two most recent CFDs
  - Link to MRs
  - Link to authors/experience
  - Relevant metrics: LOC, coverage, . . .
- Checklist of suggested actions

Example: Risky File Author View

candidate risky file list

Format 1 - Example of Login Page

Sorted by Number of Deltas User submitted for this file



Authors of [redacted]erge-ui/application[redacted]i/WindowLaunchPad.xaml.cs

Login	Name	email	Phone	NDelta(this file)	TotDelta (allfiles)	From(allfiles)	To(allfiles)
[redacted]ro	[redacted]	[redacted]@na-west.exchange.avaya.com	+1 908 [redacted]	642	34544	2007-07-26	2013-04-12
a[redacted]r				351	10820	2008-07-22	2011-10-21
k[redacted]ia	A[redacted]i	[redacted]@apac.exchange.avaya.com	+1 408 [redacted]	225	111795	2007-08-14	2013-04-12
[redacted]				194	81407	2007-01-18	2011-12-20
[redacted]ov1	D[redacted]yanov	[redacted]@mera.ru	+7 831[redacted]	121	16799	2012-02-03	2013-04-01
s[redacted]nov	Se[redacted]nov	[redacted]@mera.ru		115	4176	2010-09-05	2013-04-16
m[redacted]i				88	4858	2010-08-09	2013-01-15
jm[redacted]n	Jo[redacted]n	[redacted]n@na-west.exchange.avaya.com	+1 303 [redacted]	70	910	2011-11-10	2013-03-27
ar[redacted]ar				68	236143	2010-09-03	2012-07-19

⋮

Link to Post Entry of Login;
No link == No longer with Avaya

Information about the user across all file in all the repositories scanned

Expert assignment and training

- Use file authorship to determine/assign SME
- SME is trained how to use the report and checklist
- SME examines the report to:
  - Determine action for each risky file
  - Schedule the action

SME Recommendations

- No action required if
  - E.g., will become unused; just changed with a risky file
- Control if
  - E.g., little active development in the future
- Control examples
  - Extra review SME+Owner, and testing for any change
  - If many authors: create a brief design/test guide
- Restructure if
  - Development in the future *and* the file is too fragile
- If no remaining authors: assign a file owner

Update on status

- Created candidate sets of risky files for 45 projects.
- Held training sessions with 17 of these projects
- 7 of these projects are defining actions

Discussion

- Use of Big Data
  - To make quality visible to multiple stakeholders
- Enable SMEs to take action
  - By (usually) justifying their intuition
  - By providing quantitative evidence for management

Discussion

- A patchwork on cutting-edge techniques
  - Data mining
  - Risk prediction
  - Expertise browser (link code and people)
  - Relationship among files in different repositories
- Feedback from early users
  - Need to show or drill-down to detail: code, MRs, people
  - Multiple forms of presentation
  - Role-specific aggregation
  - Bug in another project: DILLIC/DILLIGAD?