luction	Literature Survey	Do Time Problems Exist?	

Escaping the Time Pit: Pitfalls and Guidelines for Using Time-based Git Data



Jigyasa Chauhan Samuel W. Flint Robert Dver

> Department of Computer Science and Engineering University of Nebraska-Lincoln

2021 Mining Software Repositories Conference

S. Flint, J. Chauhan & R. Dyer

Time Data is Everywhere!

Almost any repository data we touch has a time component:

- Commits
- Issues
- Pull Requests
- Forge metadata
- Logs
- Etc.

But like any data, we need to be careful handling it.



(https://latestaesthetics.com/products/melting-dali-clock)

イロト イヨト イヨト

э

Research Questions

- 1 How many MSR papers rely on time-based data?
- What kinds of data include time?
- 3 What filtering or cleaning techniques are used?
- 4 Is bad time-based Git data common?

イロト イヨト イヨト

I nar

Selection Criteria

- Mining Software Repositories (MSR) conference
- Technical track & Data Showcase papers
- 2004–2020
- 580 technical + 110 data showcase = 690 papers
- Filtered by time-related keywords, e.g., minute, hour, epoch, interval

	Do Time Problems Exist?	
0000		

Selection Criteria

- 282 matched the keyword search
- 43 papers were removed after manual inspection
- RQ1 result: 239 papers (35% of 690) were retained

Introduction 00		Literature Survey 00●00		Do Time Problems E 00	xist?	Results 000
6						
Survey RQ2 result:	Kesults Kinds of Data					
O	Search or jump to	7 Pull requests Issues	Marketplace Explo	re	Ċ +• ()·	
🛱 ro	obert-strandh / SICL			⊙ Unwatch → 72	* Unstar 771 * Fork 60)
<>	Code 🕕 Issues 24	ੀ Pull requests 3 ⊙ Actions	III Projects III		orge Metadata)	
29	master - 31%	7% ⊳otags	Go to file Add file	· ✓ Code -	About	
0	robert-strandh Take into acc	count call instructions to compute 2	c8c776 44 minutes ago	• 22,192 commits	A fresh implementation of Common Lisp	
	Code	614 to (a Vous Sa) instructions to	compute the new input.	44 minutes ago	🛱 Readme	
	Grammar	Shadow names of symbols OR, *, a	ınd +.	2 months ago	ملِغ View license	
	Ideas	Initial commit.		11 years ago		

Make the text version of the abstract relevant for this paper.

Im S. Flint, J. Chauhan & R. Dyer

Papers

Releases

4 days ago

	Do Time Problems Exist?
00000	

Survey Results Data Sources



S. Flint, J. Chauhan & R. Dyer

	Do Time Problems Exist?	
00000		

Survey Results RQ3 result: Filtering and Cleaning Techniques

- None explicitly mentioned (146, 62%)
- Selecting within a time window (24, 10%)
- Describing a cut-off date (15, 6%)
- Using a Custom Condition (7, 3%)
- Coalescing Changesets (5, 2%)
- Using an Analysis Tool (4, 2%)
- Correcting Date Formats (3, 1%)

 Introduction
 Literature Survey
 Do Time Problems Exist?
 Results

 00
 00000
 00
 0000

Survey Results RQ3 result: Filtering and Cleaning Techniques

- None explicitly mentioned (146, 62%)
- Selecting within a time window (24, 10%)
- Describing a cut-off date (15, 6%)
- Using a Custom Condition (7, 3%)
- Coalescing Changesets (5, 2%)
- Using an Analysis Tool (4, 2%)
- Correcting Date Formats (3, 1%)

イロト 不得 トイヨト イヨト

= nan

	Do Time Problems Exist? ●0	

Studying Git

- GitHub is the most common data source
- VCS data is the most studied kind of data
- What would a time problem look like? How frequently are they found?

	00	



◆□▶ ◆□▶ ◆臣▶ ◆臣▶ 臣 の�?

S. Flint, J. Chauhan & R. Dyer

Repartment of Computer Science and EngineeringUniversity of Nebraska–I incoln

	00	



▲□▶ ▲□▶ ▲目▶ ▲目▶ 目 のへぐ

S. Flint, J. Chauhan & R. Dyer

Repartment of Computer Science and EngineeringUniversity of Nebraska–I incoln

	00	



◆□ > ◆□ > ◆臣 > ◆臣 > ○ ■ ○ ○ ○ ○

S. Flint, J. Chauhan & R. Dyer

repartment of Computer Science and EngineeringUniversity of Nebraska–Lincoln

	00	



▲□▶ ▲圖▶ ▲≣▶ ▲≣▶ ■ めんの

S. Flint, J. Chauhan & R. Dyer

repartment of Computer Science and EngineeringUniversity of Nebraska–Lincoln

Observed Dataset Problems

- From Boa's 2019 GitHub Large Dataset
- 26,252 suspicious due to being out of order, from 4,744 projects
- 4,735 suspicious due to being too old, from 82 projects
- 11 suspicious due to being in the "future", from 3 projects

Introduction	Do Time Problems Exist?	
		000

Proposed Mitigation

- Filter by Project
- Filter by Date
- Remove projects modified by git-svn

◆□ > ◆□ > ◆三 > ◆三 > 三 ・ ○ < ♡

S. Flint, J. Chauhan & R. Dyer

repartment of Computer Science and EngineeringUniversity of Nebraska–Lincoln

Conclusion

Literature

- RQ1: At least 35% of MSR studies consider time-based data
- RQ2: VCS is the most common, most often from GitHub
- RQ3: Filtering & cleaning of time-based data isn't commonly described

RQ4: Mining GitHub

- Out-of-order commits most common error, 26,252 from 4,744 projects
- For "too old" commits, git-svn seems connected
- Causes likely due to tools, misconfiguration or user error

Replication Package: https://doi.org/10.5281/zenodo.4625288