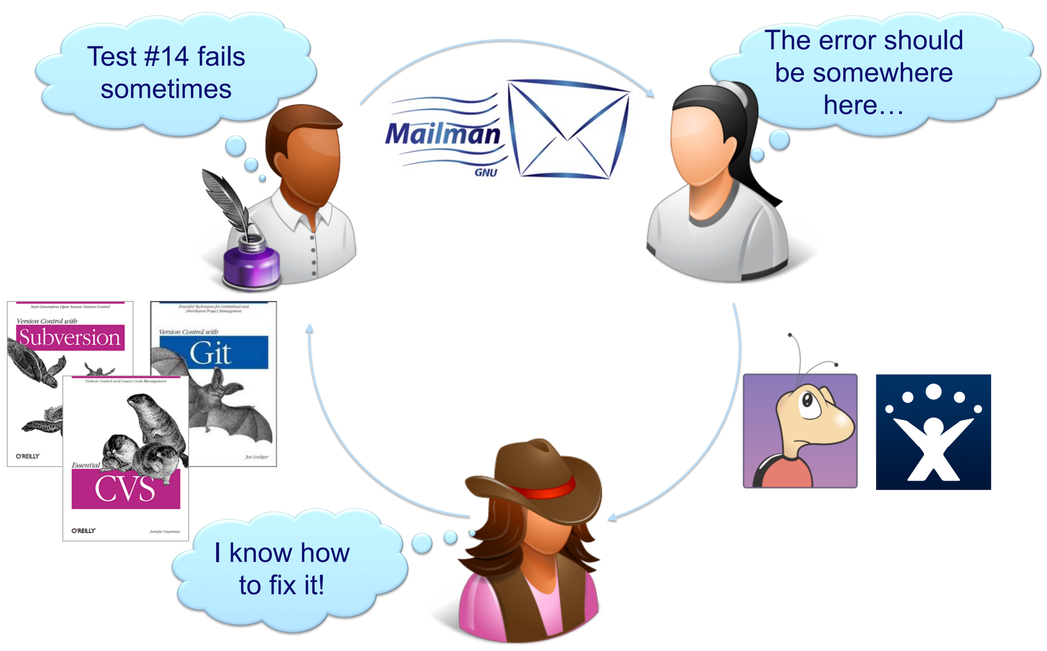
Who's who (in GNOME)?

Erik Kouters ◆ Bogdan Vasilescu ◆ Alexander Serebrenik ◆ Mark G.J. van den Brand
e.t.m.kouters@student.tue.nl ◆ b.n.vasilescu@tue.nl ◆ a.serebrenik@tue.nl ◆ m.g.j.v.d.brand@tue.nl
Eindhoven University of Technology, The Netherlands

Software ecosystems

- different projects
- different repositories per project (version control systems, mailing lists, …)
- same community of developers



Multiple aliases

Contributors sign off using a `<name, email>` alias.
The same person often uses multiple aliases.

Names:
- John Travolta
- Travolta John
- J Travolta
- John
- John Trabolta
- John J. Travolta
- John Joseph Travolta
- John "Bone" Travolta

Emails:
- j.travolta@domainA
- john.travolta@domainB
- john DOT travolta AT domainC
- jtravolta@domainD
- john@domainE

Identity merge algorithms

Merge aliases belonging to the same "real" person.

- Goeminne & Mens (2011) – simple, state of the art

```
<John Travolta, j.travolta@domainA>
<John Travolta, john@domainB>
```
✓

```
<John Travolta,    j.travolta@domainA>
<John J. Travolta, john@domainB>
```
✗

```
<John Travolta,   john@domainA>
<John F. Kennedy, john@domainB>
```
✗

- Bird et al. (2006) – more complex, inspiring

```
<John Travolta, bone@domainA>
<John,          jtravolta@domainB>
```
✓

```
<John Travolta, john@domainA>
<Joey Travolta, joey@domainB>
John ~Levenshtein Joey
```
✗

```
<John Travolta, john@domainA>
<Travolta John, travolta@domainB>
John !~Levenshtein Travolta
```
✗

Latent Semantic Analysis

```
<John Travolta,          johnt@domainA>
<John Joseph Travolta,   johnt@domainA>
```

```
johnt@domainA:
 {john, johnt,
 joseph, travolta}
```

Document-term matrix

	johnt@domainA	j.travolta@domainB	..	..
john	1	..	..	..
johnt	1	..	..	..
joseph	1	..	..	..
jtravolta	8/9	..	..	..
travolta	1	..	..	..

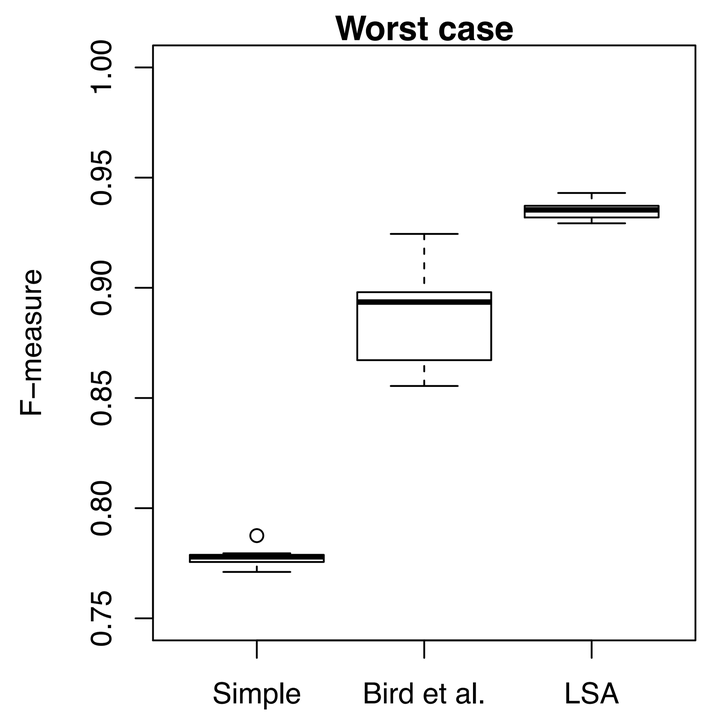
- Singular value decomposition
- Rank (noise) reduction
- Cosine between documents

Merge documents with high similarity!

max similarity(jtravolta, {john, johnt, joseph, travolta})
= similarity(jtravolta, travolta) = 1 – Levenshtein(jtravolta, travolta)
= 1 – 1/9 = 8/9

Results for GNOME

- 8618 different aliases
- only 4989 unique!



- LSA outperforms existing algorithms