FLOSS 2013: A Survey Dataset about Free Software Contributors: Challenges for Curating, Sharing, and Combining

Gregorio Robles, Laura Arjona-Reina, Alexander Serebrenik, Bogdan Vasilescu, Jesus M. Gonzalez-Barahona

Universidad Rey Juan Carlos, Universidad Politecnica de Madrid, TU Eindhoven (The Netherlands)
grex@gsyc.urjc.es, laura.arjona@upm.es, {a.serebrenik,b.n.vasilescu}@tu-eindhoven.nl, jgb@gsyc.es

http://floss2013.libresoft.es

Goals

The goals of this MSR data paper are following:

1. **Curated data:** To offer a curated data set with data from over 2,000 FLOSS contributors.
2. **Combination of data:** To present a case study, the challenges and issues of an "augmented" use of the data together with public data from other sources [5].

The complete questionnaire, including answers, can be obtained from `http://floss2013.libresoft.es/downloads/questions/FLOSSSurvey2013_en.pdf`.

Relevance

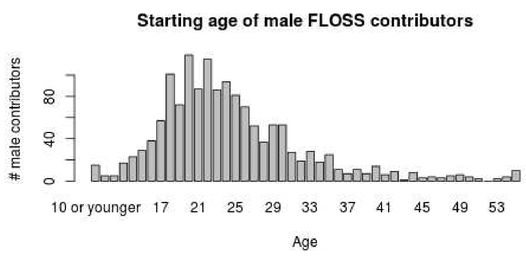
- Data obtained by means of surveys with research purposes is seldom shared.
- One of the reasons for the lack of sharing is that these data sets contain private data or personally identifiable information.
- However, much of the information obtained by means of a survey is very difficult (if not impossible) to obtain by other means.
- Linking data obtained from surveys with other data, gathered by traditional mining software repositories means, may provide new insights and allow for further discoveries.

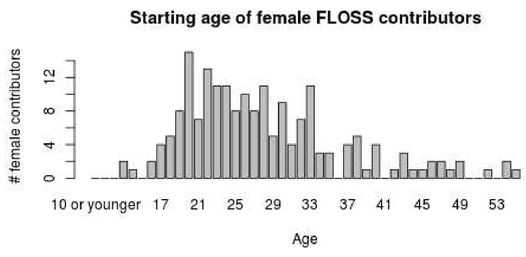
Methodology

The survey methodology of the FLOSS 2013 survey has been the same as the one of the original FLOSS survey: an open web-based survey, where participants are self-selected.
The 58 questions can be classified into following areas:

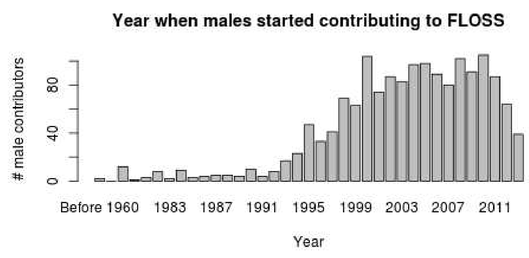
- **Personal situation** (gender, civil status, number of children, country of birth and of residence/work)
- **Education** (highest level of education, level of English)
- **Professional situation** (profession, satisfaction, income)
- **FLOSS perspective** (free software vs open source)
- **Development** (age when joining FLOSS, reasons and motivations for joining, reasons and motivations for still participating, earn money with FLOSS)
- **Technology** (favorite editor, programming languages)
- **Economic and community rewards** (job opportunities, expectations from other developers, challenges)

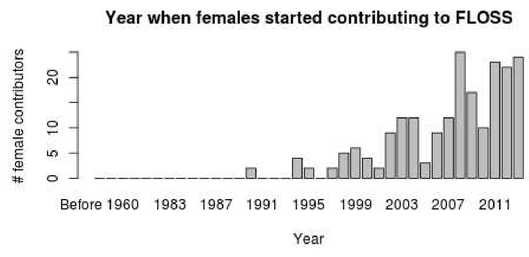
Some Results (I)





Some Results (and II)





Combining data and Privacy

- Combining data provides with an *augmented* view of the matter of study.
- Sometimes some demographic variables affect the results of an investigation, as it is well known from other fields of research; for instance, gender [6].
- Sharing and combining data supposes several risks. Traditional anonymization techniques have proven to be limited [3]. Legal, ethical and practical issues have to be considered.
- Survey respondents should not be recognized from the output if personal data is shared.
- There are several approaches that have been proposed to ensure *secure* anonymization and that we would like to study in the next future:
  1. k-anonymity [4]
  2. l-diversity [2].
  3. t-closeness [1].
- In the meantime, we have combined the data internally.

Combining data: Case study

- Our aim is to link the FLOSS survey data with data from other sources, in this case data from StackOverflow, to show its potential uses.
- StackOverflow is the largest Q&A website for programmers, with more than 2.3M users registered as of September 2013.
- To automatically infer gender for StackOverflow users, we used a previously-validated [5] name-based gender resolution tool. The tool (`https://github.com/tue-mdse/genderComputer`) tries to infer a person's gender based on their name and, if available, their location.
- The samples are composed of 1,476 FLOSS survey respondents that provide a complete and valid (at least from its construction) e-mail address, which has been hashed with MD5. For StackOverflow, we have 2,091,063 distinct MD5 hashes of e-mail addresses out of a total 2,332,406 total MD5 hashes gathered.s

Combining data: Case study Results

- As a result of matching the MD5 hashes in both datasets, we have obtained 451 matches. From these, 439 had provided gender information in the FLOSS survey.
- Considering the gender resolution algorithm used with StackOverflow, we have identified 227 correct gender matches.

Table 2: Measures when combining StackOverflow gender resolution results with the FLOSS 2013 survey.

Gender	Male	Female	Total
Precision	0.97	0.55	0.90
Recall	0.54	0.39	0.52
F-measure	0.69	0.46	0.66
MCC	0.26	0.42	0.62

- Values of MCC are between -1 and +1, representing +1 a perfect prediction and 0 no better than a random prediction. The formula of the MCC is:

$$MCC = \frac{t_p * t_n - f_p * f_n}{\sqrt{(t_p+f_p)(t_p+f_n)(t_n+f_p)(t_n+f_n)}}$$

Acknowledgments & References

The work of G. Robles and J. M. González-Barahona has been funded in part by the Spanish Government under project Sobre-Sale (TIN2011-28110). The research of B. Vasilescu has been financed by the Dutch Science Foundation with project NWO 600.065.120.10N235.

References

[1] N. Li, T. Li, and S. Venkatasubramanian. t-closeness: Privacy beyond k-anonymity and l-diversity. In *ICDE*, volume 7, pages 106–115, 2007.

[2] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam. l-diversity: Privacy beyond k-anonymity. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1):3, 2007.

[3] P. Ohm. Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review*, 57(6), 2010.

[4] L. Sweeney. k-anonymity: A model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(05):557–570, 2002.

[5] B. Vasilescu, A. Capiluppi, and A. Serebrenik. Gender, representation and online participation: A quantitative study. *Interacting with Computers*, page iwt047, 2013.

[6] H.-Y. Wang and Y.-S. Wang. Gender differences in the perception and acceptance of online games. *British Journal of Educational Tech*, 39(5):787–806, 2008.