Building Knowledge in Open Source Software Research in Six Years of Conferences

Fabio Mulazzani, Bruno Rossi, Barbara Russo, and Maximilian Steff

Center for Applied Software Engineering (CASE),
Free University of Bozen-Bolzano,
Piazza Domenicani, 3, 39100 Bolzano, Italy
{fmulazzani,brrossi,brusso,maximilian.steff}@unibz.it

Abstract. Since its origins, the diffusion of the OSS phenomenon and the information about it has been entrusted to the Internet and its virtual communities of developers. This public mass of data has attracted the interest of researchers and practitioners aiming at formalizing it into a body of knowledge. To this aim, in 2005, a new series of conferences on OSS started to collect and convey OSS knowledge to the research and industrial community. Our work mines articles of the OSS conference series to understand the process of knowledge grounding and the community surrounding it. As such, we propose a semi-automated approach for a systematic mapping study on these articles. We automatically build a map of cross-citations among all the papers of the conferences and then we manually inspect the resulting clusters to identify knowledge building blocks and their mutual relationships. We found that industry-related, quality assurance, and empirical studies often originate or maintain new streams of research.

Keywords: Systematic Mapping Study, Cross-citations.

1 Introduction

Since its origins, the diffusion of the OSS phenomenon and the information about it has been entrusted to the Internet and its virtual communities of developers. As such, information on OSS has grown exponentially resulting in a vast quantity of data readily available. This data has attracted the interest of researchers and practitioners aiming at formalizing this information into a body of knowledge (e.g. [12], [13]). For this reason, in 2005, a new series of conferences on OSS (OSS conference series[1]) and, in 2009, a new journal[2] have been established. These initiatives have substantially contributed to initiate a long process to ground knowledge in OSS that masters data from different sources and consolidates them into well-accepted concepts and their mutual relations. As such, they represent a valuable source for understanding how OSS knowledge has been created and how it will evolve in the future. Our work tackles this issue proposing a systematic mapping study of the papers of the OSS conference series.

[1] International Symposium on Open Source Software initiated in 2005, in Genoa, Italy.
[2] International Journal of Open Source Software and Processes, http://www.igi-global.com

S.A. Hissam et al. (Eds.): OSS 2011, IFIP AICT 365, pp. 123–141, 2011.

Systematic Mapping Studies (SMS) and Systematic Literature Reviews (SRL) are techniques of knowledge synthesis. These techniques are typically based on manual inspections of articles ([23], [24], [25], and [37]). Manual inspection requires significant effort in mining large sets of articles. On the other hand, a complete automated inspection can produce inaccurate results. In our work, we propose a semi-automated approach to mine articles' repositories for systematic mapping studies. We automatically inspect articles to build a map of cross-citations and then we manually inspect the resulting clusters to identify the building blocks of knowledge in OSS and the social network of the community maintaining it. We selected the complete database of articles of the OSS conference series since its origin in 2005[3]. Our choice is driven by three criteria: 1) papers are peer reviewed - this excludes for example the MIT repository[4], Apache conferences[5], OpenOffice.org conferences[6], etc…; 2) the series' mission is to disseminate knowledge in OSS - this excludes traditional journals in software engineering; and 3) papers report more than a group discussion - this excludes workshops or one day events co-located with larger non OSS events. The result of this study aims at answering two major questions:

RQ1. Is there any social network underlying the research production at the OSS conference series?

RQ2. What are the major streams of research proposed at the OSS conference series?

Our answer to RQ1 will identify the cornerstone papers and the links among them across the years. Links will express the relation among the authors by means of the connection of their research production. The analysis will reveal unexpected and undeclared connections among authors as well as lack of connections among conceptually related papers. This will also illustrate the self-sustainability of the OSS conference series and the value that it provides to the OSS community. In addition, using the results in [37], our research will also discuss how empirical studies fit the network. An answer to RQ2 will help build the baseline for future investigations in OSS research or to extend existing ones.

In the following section, we introduce related work and motivate our work. Section 3 presents our method of SMS and Section 4 explains our analysis methodology. In Section 5, we explore the results of our analysis and describe the clusters of papers we identified, followed by a summary of our findings in Section 6. We close with the conclusions and limitations.

2 Background and Motivation

The software engineering community has been increasingly adopting Evidence-Based Software Engineering (EBSE) approaches to build discipline-specific bodies of knowledge such as Inspection, Testing, and Requirements Engineering ([7], [39]). Apart from the traditional ways of doing literature review, also called ad-hoc reviews,

[3] https://pro.unibz.it/staff/brusso/PapersUsed.html
[4] http://opensource.mit.edu/
[5] http://na11.apachecon.com/
[6] http://www.ooocon.org

the EBSE practice uses Systematic Literature Reviews (SLRs) ([6], [24], [25], and [26]) and Systematic Mapping Studies (SMSs) ([7], [23], and [29]) as robust methodologies of searching, selecting, analyzing, and synthesizing literature and aggregate evidence on a specific topic. As such, SLRs and SMSs are called secondary studies as they aggregate research of other, so-called primary studies. SLR is "a means of evaluating and interpreting all available research relevant to a particular research question, topic area or phenomenon of interest" ([11], [26]). Research in SE has provided guidelines and lessons learned for performing SLR ([5], [6], [10], [25], and [34]). SMS is used to draw a landscape of reported research on a particular topic ([14], [23], [25], and [29]). Being less specific, SMS requires significantly less effort than SLR; however, it provides only a course-grained overview of the published literature. An SMS can also serve as a preparation activity before doing an SLR. SLR and SMS in OSS has been typically used to provide evidence of practices and methods for a more general SE research purpose Only recently, secondary studies have been published to investigate specific areas of OSS. In 2010, Hauge et al. performed an SLR of research on OSS adoption [16]. In 2009, Stol et al. ([36], [37]) presented an SLR on empirical papers published at the OSS conference series. To our knowledge, the first secondary study that investigated OSS as a holistic phenomenon is the work in [2]. In their work, the authors have published taxonomy of OSS mining 623 journal papers - excluding conference papers, though. In this context, our work provides an SMS on OSS as a holistic phenomenon. Our work is complementary to the work in [2] in terms of papers investigated, method of analysis, and research goal.

3 Research Method

Our method follows the concepts of a systematic mapping study [29]. In the introduction, we illustrate our search criteria for inclusion and exclusion. Following them, we select all the papers of the OSS conference series. In this section, we describe how we identify and apply classification criteria on the selected papers. Classification categories are taken from the Calls for Papers of the OSS conference series and are used to label cross-citation clusters. In particular, we automate papers classification to reduce effort of articles' inspection and enable future replications. Following [3], to increase the transparency of our method, we also detail the tools we used and the process we follow.

Worth noticing that this approach differs from the one proposed by Kitchenham in 2010 ([23]). Kitchenham uses citations to identify most and least cited papers. We propose here to extend this approach using citations to determine streams of research.

3.1 Creating the Directed Graph of Cross-Citations

We collected the PDF files of papers from the Springer repository. We developed an application (PDF Analyzer[7]) that (a) converts the papers from the PDF format to a textual representation, and then (b) injects this representation into an XML file with nodes corresponding to the paper's sections. The application allows user intervention during the conversion process. We have also sampled part of the XML files to verify

[7] Apache PDFBOX library to convert PDF to TXT and DOM4J library to create XML the file.

and validate the tool output. When problems occurred, we manually corrected them with the aid of the original PDF file.

We create a Python script that 1) extracts all papers' titles and conference years from the XML files, 2) parses all the references for all the paper titles, 3) for every hit, extracts conference names, and 4) goes over the references again to identify possibly missing titles using the different variations of conferences' names.

We noticed that conferences' titles significantly vary in that we identified 17 different variations. At the end, we also manually further checked for missed citations. The final output of the Python tool classifies papers by year of conference and by citations and passes them to GraphViz[8] to display the final graph (Fig. 11).

3.2 Descriptive Analysis of Cross-Citations

Before performing any inspection of the clusters, we have analyzed the citations of the papers we found. Table 1 shows the distribution of citations over the years. Articles refer to full and short papers if any.

Table 1. Number of articles cited by or citing another article of the OSS conference series

	2005	2006	2007	2008	2009	2010
Cited	25	15	16	15	8	-
Citing	-	12	13	17	11	13
Total	83	41	51	42	28	40
Isolated	58	23	33	21	16	27
% citing/cited	31%	44%	35%	50%	43%	32%
# articles cited >2	11	3	4	3	0	0

In particular, Table 1 shows that there are a good number of *isolated articles* in that they do not cite other papers.

3.3 Inspecting the Graph

Fig. 11 illustrates the complete directed graph of articles that cite or are cited by other articles. Each article is a node and citations are edges. Articles-nodes mapping is provided at https://pro.unibz.it/staff/brusso/PapersUsed.html. We define *fan in* as the number of citations to a paper and *fan out* as the number of citations from a paper. We define the *distributor of knowledge* as a node in the graph that has at least fan-in equals two and an *attractor of knowledge* a as node with at least fan-out equals two. A pure distributor has zero *fan out* and a pure attractor has zero *fan in*. A node can also simultaneously be a distributor and an attractor. A node that is a distributor with fan out one is a *router* as it branches the knowledge of the single citation into different articles. For example, node #107 is a router that distributes the knowledge of paper #82 to five other articles (Fig. 2). Fig. 1 displays the types of nodes. A path is a set of nodes connected by edges following the direction of the graph.

[8] http://www.graphviz.org/

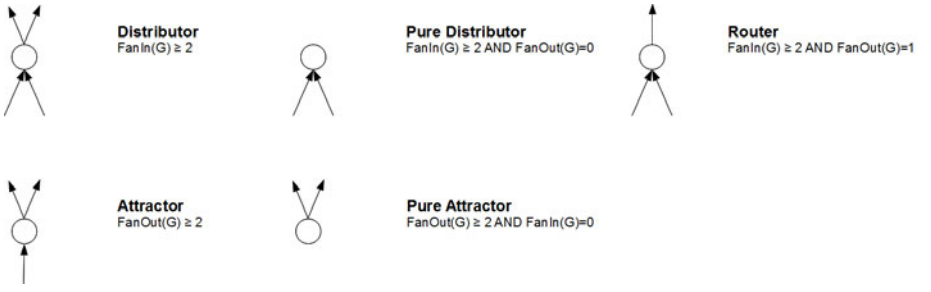

Fig. 1. Types of nodes

To determine a research area in OSS, we make three assumptions:

i) Pure distributors determine research areas;
ii) A path originating from a pure distributor and leading to a pure attractor or a dead end (a paper that has only one fan out) determines an area of research in OSS;
iii) Paths starting from a pure distributor determine a cluster.

Thus, we start from a pure distributor, for example #82, and then follow one of its links. We follow links in opposite direction downwards in the graph until we reach a pure attractor or a dead end. Then we aggregate all the paths from a pure distributor to define a cluster. Finally, we add all the dead ends cited by an attractor of the cluster. For example, the edge linking #82 to #107 determines an area of research originated from paper #82 including the dead ends node #41 and #81. The five paths originated by paper #82 determine a cluster.

To label paths, we have used the taxonomy of the Call for Papers (CfP) of the OSS conference series[9]. Typically, a CfP includes the major topics of the conference. Consequently, accepted papers concern topics listed in the CfP. We have also considered the classification proposed in [37]. Unfortunately, the classification was too high-level for our analysis.

We have identified twenty - one clusters in the graph (Fig. 11). Four are bipoles - clusters of two articles - and two are isolated clusters. Eleven in 2005, two in 2006, and two in 2007 originate the fifteen clusters. Thirty-four are empirical papers according to [37]. Note that the majority of the pure distributors that originate the largest clusters are empirical. Table 2 lists pure distributors and attractors that define the largest clusters - papers with more than three fan-out or fan-in. It also classifies them as empirical according to [37].

To identify the reason of the citation, we have manually looked up each citation in the text. We have used the wording of the authors and the position of the citation within the article structure to understand the reason for each citation. If, for example, a citation is located in the “Background” section only and it is given to justify the work, then we label the corresponding link as motivation of work.

[9] http://ossconf.org/

Table 2. Major Distributors, Attractors, and Routers

Distributors and routers
2005
#5 pure distributor and empirical paper ([1]), #8 pure distributor and empirical paper ([30]), #17 pure distributor and empirical paper ([27]), #44 pure distributor ([31]), #82 pure distributor ([22])
2006
#84 pure distributor and empirical paper ([39]), #107 router and empirical paper ([21]), #119 router ([8]), #121 pure distributor and empirical paper ([32])
2007
#127 pure distributor ([38]), #128 pure distributor and empirical paper ([14]), #138 router and empirical paper ([20])
2008
#180 distributor/attractor and empirical paper ([19])
2009
#234 router ([9])
Attractors
2008
#175 Platform for research ([15])
2009
#230 Framework ([33]), #233 Extensive background ([18])
2010
#305 Includes SLR ([17]), #312 Framework ([35])

4 Classification of the Articles

We have read all the articles following the patterns defined by the clusters. The reading confirms that each path originated from a pure distributor determines a well-defined perspective of research that takes its motivation from the distributor. In many cases, we are also able to identify authors that contributed the most to a given research area and determine the semantic of the cross-citations besides the motivation of work. In the following, we report of this analysis per cluster. The number of the pure distributor names clusters.

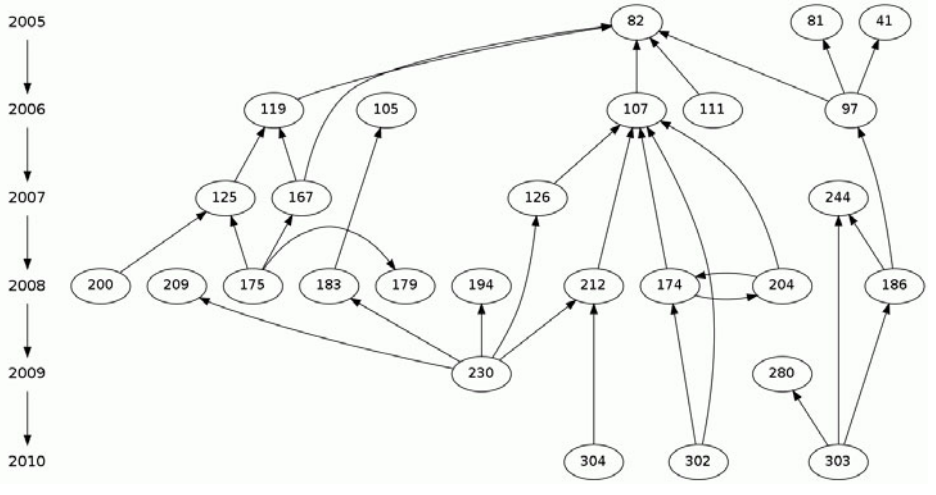

Fig. 2. Citations cluster originated from paper #82

Cluster #82. The largest cluster originates from node #82. Paper #82 introduces the OSSmole project (later called FLOSSmole). OSSmole is a repository of data, scripts, and analysis of data collected from OSS projects. The cluster is then branched into five links (Fig. 2). Three links identify three major sub-clusters, defined by router #119, router #107, and node #97. The sub-cluster defined by router #107 is generally concerned with developer communities and social network analysis of these communities. Links are rather strong. Paper #174 cites paper #107 to motivate its research and the use of a metric (outdegree centrality), and cites paper #204 as example of method of analysis. Two of the articles in this cluster focus on the same project (Apache).

Howison and Crowston have been the major contributors. The branch ends in 2010 with the work of Conaldi and Rullani that proposes a global perspective of F/OSS network structure mining the SourceForge repository. Paper #119 on the future of OSS data mining starts a new branch that focuses on analyses and improvements of project mining tools. In this sub-cluster, we found citations motivated by the use of the same repository or the same research goals. The branch ends in 2008 with recommendations for the design of research infrastructure in OSS by Gasser and Scacchi (paper #175). Paper #97 originates a research branch on the analysis of code artefacts for modelling maintenance processes and specializes over the years in bug fixing processes. Although article #97 cites #82, it does not use OSSmole directly mining CVS log files. The major contributors in this branch are Dalle and den Besten and the majority of the articles focus on the Firefox community.

Cluster #44. Paper #44 introduces to practices for quality assurance in OSS projects (Fig. 3). Paper #44 generates three sub-clusters on OSS deployment, project's activity, and community building and participation. Article #156 connects the last two topics by means of increase of the community size and their activity growth. In cluster #44, we were not able to identify major contributors as different authors contributed to the research streams and all the citations were to motivate the work.

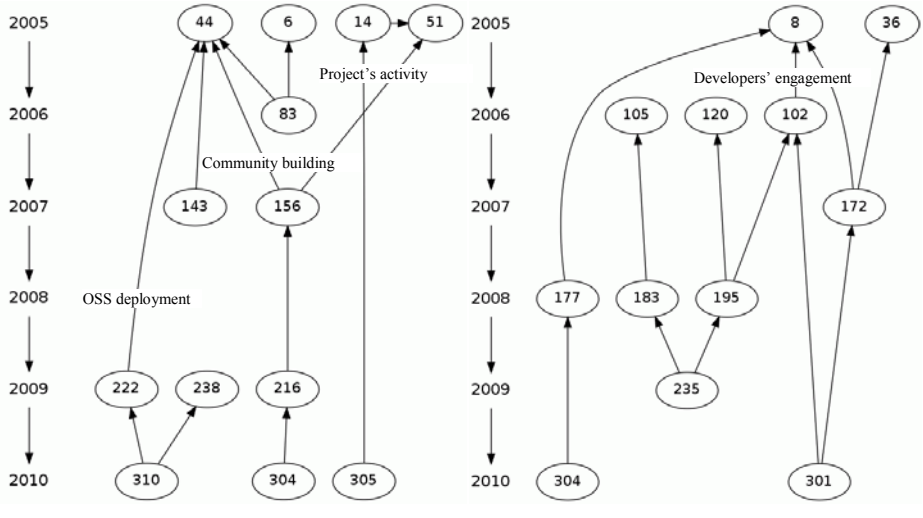

Fig. 1. Citations cluster originated from papers #44 and #8

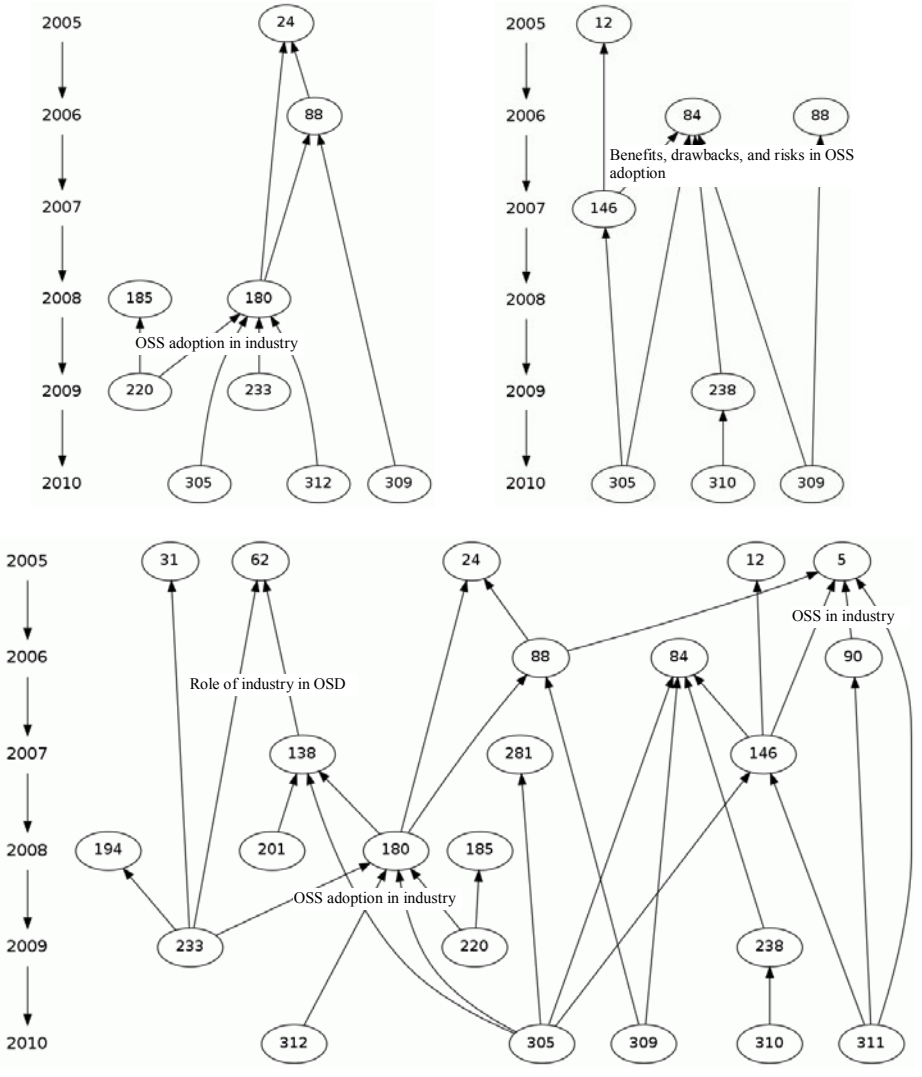

Fig. 4. The four clusters #5, #24, #62, and #84

Cluster #8. Paper #8 concerns the involvement of volunteers in OS projects (Fig. 3). A first branch defined by router #102 evolves into the topic on developers' engagement in OS projects. Developers' engagement is used to correlate OSS to agile methods (#195), to evaluate the involvement of companies (#172), or to investigate communication among developers (#235). In the majority of the subtopics, Barahona and Robles and Capiluppi and Adams have been the major contributors. Citations define rather strong connections among papers of this cluster. For example, articles #102, #120, #195, #235 use data from KDE, whereas papers #304 and #301 use the same metric of paper #177 and #102 respectively. In some of the cases, they share some of the authors, too.

Cluster #5. Cluster #5 is a big cluster connected to three other sub-clusters (#24, #84, and #62 in Fig. 4). These clusters focus on the relation between OSS and industry. The empirical pure distributor #5 summarizes the works of an international workshop on pros and cons of the use of OSS in industry. This paper spins off into two contributions of people that participated in the workshop (#90) and of some of its authors (#146 and #88). Looking at the affiliations of the authors, we call cluster #24 the industrial "Scandinavian case": it starts with a case of interest in OSS in the Finnish industry (#24), then includes the Swedish one in 2006 (#88) and closes with the Norwegian one in 2008 (#180). From paper #180, Hauge, Sorensen and Conradi initiate a new research theme on OSS adoption in industry (#233, #220, #309, #312, and #305). Paper #62 uses the Capability Maturity Model to assess the migration from closed to open software development in industry. This paper initiates a research investigation on the role of industry in Open Source Development (OSD) (router #138) and in particular, the migration of an existing business model to an open source one (#233). It further evolves in OSS adoption in industry (#180). Cluster #84 originates from the paper of Ven and Verelst (#84) and gives impulse to the specific perspective of OSS adoption in industry that relates to benefits, drawbacks, and risks. The authors that contribute the most to cluster #84 are Hauge and Conradi. The majority of the citations of the four clusters concerns motivation of work, but there are links that connect papers by the same method of analysis (e.g. #90 and #311). Worth noticing is that all the three papers of 2006 (#84, #88, #90) report of a case study in industry, but only the first two are as empirical according to [37].

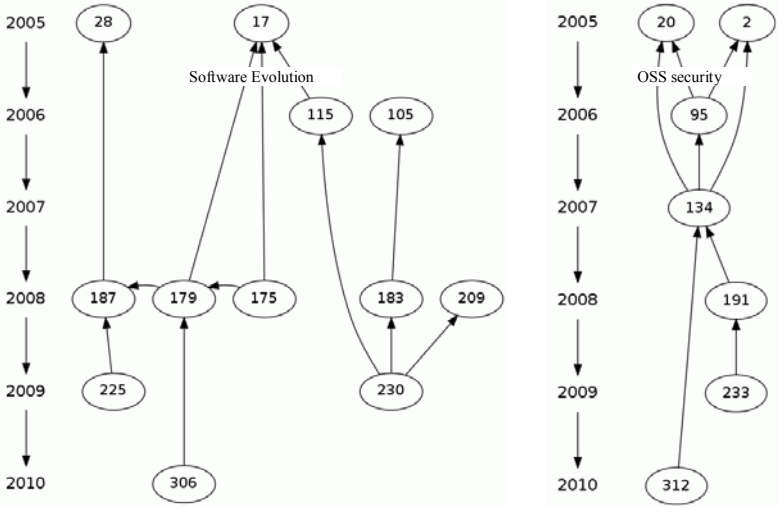

Fig. 5. Clusters #17 and #20_2

Cluster #17. Cluster #17 is originated by the work on software evolution by Koch (Fig. 5). The research has developed into software evolution as total growth of software (Riehle et al., #179, #187, and #225) and evolution of OS communities as measure of their governance (#230). An independent sub-cluster (originated by paper

#28) is connected with the software evolution theme as it deals with the analysis of commits in distributed development. The link is provided by a chain of citations in 2008 that connects the work of Gasser and Scacchi (#175) on a research framework for multi-disciplinary studies in OSS (including the "laws of software evolution") with the study of continuous integration in OS development as an example of mining multiple data sources (#187). In this cluster, we identify a citation from the paper of Sirkkala et al. (#230) that have used part of the approach in Weiss et al. (#115) and a citation from Deshpande and Riehle (#179) that uses the results of another work of them (#187) to validate their conclusions on total growth of software.

Cluster #20 and #2. Two pure distributors (#2 and #20) originate this cluster (Fig. 5). We call it the "Italian case" as the major contributors are Ardagna, Damiani, Frati and other Italians. In their three works, they deal with security of OSS. In the last two years, the area has evolved to the larger problem of selecting and integrating OSS for industry (#233 and #312) where security is especially relevant (like telecom applications).

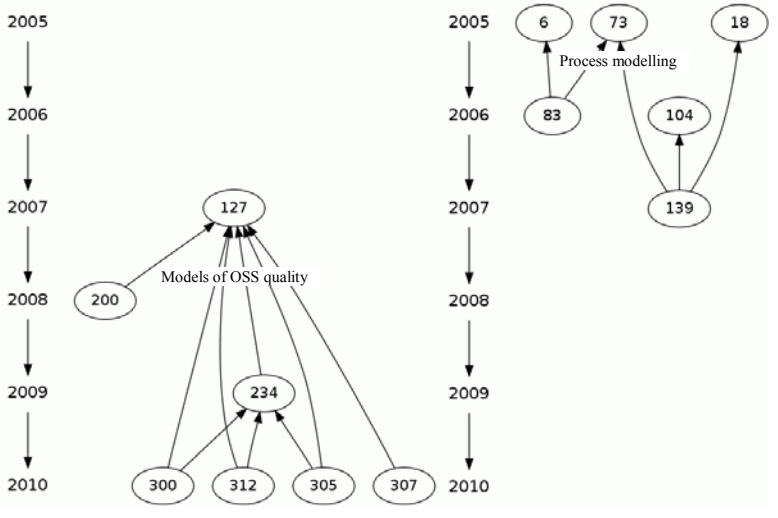

Fig. 6. Clusters #127 and #73

Cluster #127. Cluster #127 (Fig. 6) originated in 2007 with the results of the European project QualiPSo and enforced with the results on OSS trustworthiness (router #234). The majority of the works proposes and compares frameworks and comparative models of OSS quality. Major contributors are Morasca, Lavazza, and Taibi, members of the above project. In this cluster, the majority of the papers and in particular, all those in 2010 cite #127 and #234 to use the model defined there.

Cluster #73 and Cluster #121. These clusters (Fig. 6 and 7) concern process modelling and they are connected through the paper of Jensen and Scacchi in 2007 on guiding the discovery of OSS processes with a reference model (#139). Two branches

derive from the pure distributor #73: on modelling communication and information exchange in processes and on modelling the process as a whole. Paper #139 motivates its research citing the problem of managing information in distributed development (#73) and issues in creating theoretical models of OS processes (#104, #18, and #121). The research goal is organizing knowledge in OSD and providing guidance in allocating resources, selecting tools, and performing activities of OSD. Following Cluster #121, the research culminates in 2009 into an investigation of the selection of OSS products in industry as indication of reuse (#220).

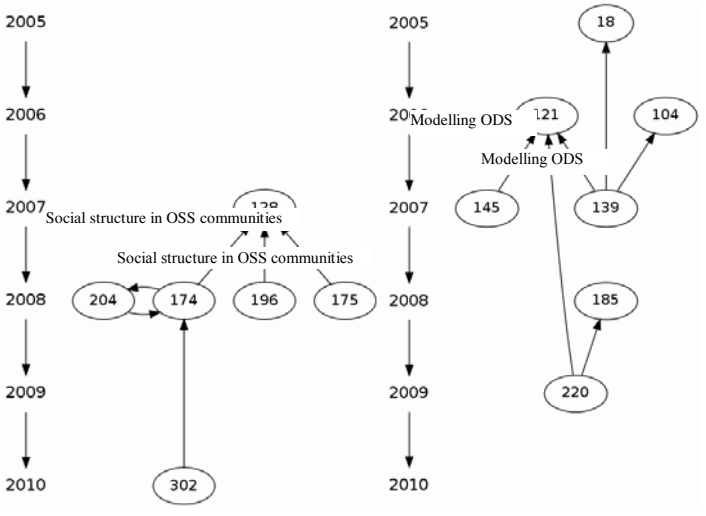

Fig. 7. Clusters #128, #121

Cluster #128. A study of network analysis on SourgeForge originates this cluster (Fig. 7). The cluster concerns social structures of OSS communities, as membership networks or communication networks. In particular, Gasser and Scacchi propose an infrastructure for research in social networks of OSS communities whereas Balieiro et al. introduce a study on the meso-level structure of OSS development as collaboration environment. Papers are authored by various members of the OSS research community and are connected by a net of citations that concerns mainly the motivation of work.

Isolated Clusters. There are six isolated clusters on the right of Fig. 11 (Fig. 8). The four bipoles concern innovation (#164 - #193), adoption (#149 - #181), services (#87 - #137) and requirements (#211 - #241). In all but the bipole on adoption, the two papers share part or all of the authors. Cluster #7 concerns measures of success of OS projects. Over the years, the concept of success evolves from a static meaning – e.g. number of hits of web searches - to an evolutionary perspective of development efficiency as the code produced over time. All the citations motivate the work and no specific author can be uniquely associated to the cluster. Cluster #68 concerns teaching OSS at university level. All the citations motivate the work.

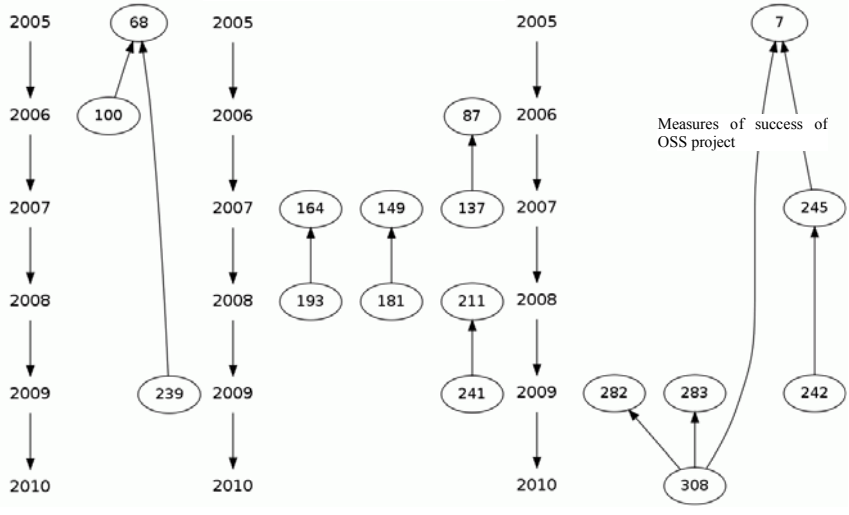

Fig. 8. Isolated Clusters

Major Pure Attractors. Pure attractors play a fundamental role to connect clusters and to summarize to some extent existing knowledge for a given purpose. As synthesizers of research, we select those pure attractors that cite more than four articles: #139, #175, #230, #233, #305, and #312 (Table 2). In #139, Jensen and Scacchi present a reference model to discover OSS processes. Paper #312 presents a framework that compares methods for evaluating OSS. Paper #305 includes a systematic literature review to identify benefits and drawbacks of OSS (Fig. 9). Paper #230 builds a conceptual framework for planning release processes of OSS. Paper

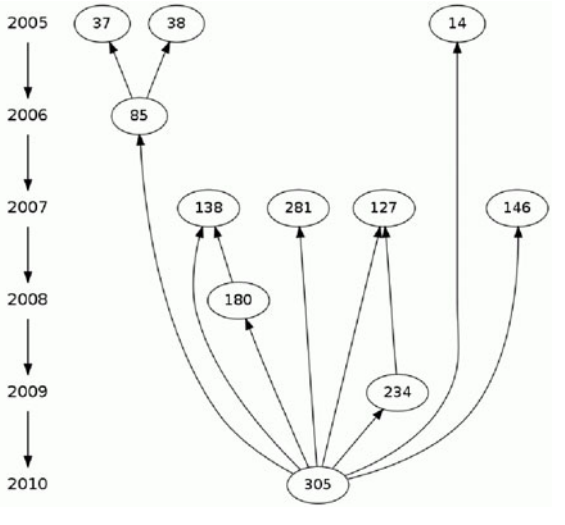

Fig. 2. Major Pure Attractor #305

#175 defines the design of a platform for investigating the OSS phenomenon through multidisciplinary research. This work also provides recommendations and critical issues in OSS research. Paper #233 includes an extensive background section that motivates the work.

4.1 Inter-cluster Connections

There are several interlinks among clusters (Figure 10). We mark a link between two clusters if there is a node that directly connects them, i.e. a pure attractor citing papers of two clusters. Connections among clusters describe mutual influence among topics and have the potential to converge to a single cluster[10]. The major findings in Figure 10 show that:

1. "Tools for data mining" is a cross cutting topic connecting various research areas, from the analysis of social structures and networks of OSS communities and their developers' engagement, software evolution, OSS security, and models of OSS quality.
2. OSS in industry specifies into "adoption in industry," "role of industry," "case studies in industry" and a stream in "benefits, risks, and drawbacks of OSS."

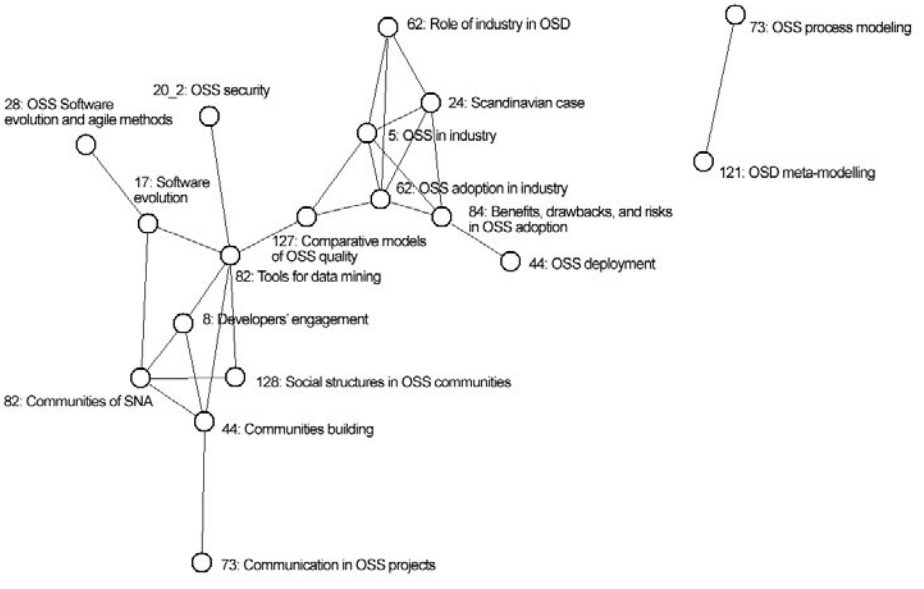

Fig. 10. Links between research streams

[10] Figure 10 does not show papers that connect clusters. This information is accessible at https://pro.unibz.it/staff/brusso/LinkingPapers.html

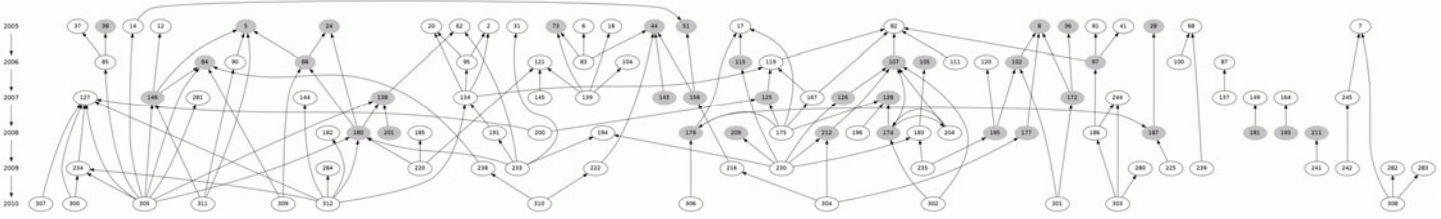

Fig. 11. The directed graph of citing articles within the OSS conference series. Grey nodes are empirical papers according to [37].

5 Discussion

Our analysis reports of well-scoped clusters of citations that highlight major areas of research in OSS and their evolution across the six years of OSS conferences. We report of a lively social network of citations (RQ1), and several streams of research established across years (RQ2).

In particular, we have found that the creation of a big repository for data mining (FLOSSmole) has originated research in social network analysis, tools for data mining, and analysis of code artefacts to understand maintenance processes, specifically bug fixing. The topic "tools for data mining" also appears to be a cross-cutting theme that connects several other research areas.

Quality assurance is another hot topic. It motivates studies on community building (as a means for QA), OSS deployment (as a factor to consider in QA), project's activity (as a measure of QA) and on comparative models (for QA).

A specific concern at the OSS conference series is the perception of OSS in industry. Research has focused on the role of industry in OSS development and after 2008 on OSS adoption. A report of a large workshop in OSS that joined researchers and practitioners of the industry in 2005 initiated this research.

A good number of papers at the OSS conference series have been dedicated to developers' engagement. Developers' engagement is used to correlate OSS methods to agile methods, to evaluate the involvement of companies, or to investigate communication among developers. This topic is also cited in works on software evolution and community building and participation.

There are themes still little cited that might have some future potential: OSS process (meta-) modelling, OSS security, Agile and OSS development methods, and teaching OSS in universities.

We have found the majority of large clusters to have a group of authors that contribute to the research stream over the years. Clusters also reveal authors' interrelations that define informal research groups and outline social network structures, like in the "Scandinavian case."

Large clusters are initiated by empirical papers with the only exception being the paper on the FLOSSmole repository.

Papers with a large number of citations are synthesizers of research often presenting a framework or a platform to guide research in OSS.

As a final remark, we want to stress that we are aware that the sample we selected does not include all the crucial papers that may have contributed to ground knowledge in OSS. Some of the relevant contributions have been published in journals not dedicated to OSS or have simply been reported on the Internet. History of knowledge synthesis is full of such examples as illustrated by the FODA paper [28], a technical report cited more than 1500 times although not peer reviewed. Our future work will extend this analysis to OSS article hubs like for example http://flosshub.org or http://pascal.case.unibz.it or the MIT repository.

We also acknowledge that fact that cross-citations are indicators of research connections, but are not unique and not necessarily the best ones. A textual similarity analysis could give finer results. Namely, we have already applied some known algorithms (e.g. cosinus, Jaro Winkler) for text similarity on the XML sections of the papers. Despite some preliminary good results on titles and authors, for longer

sections the algorithms proved to be inaccurate. We plan to use more sophisticated techniques of string similarity and a better data cleansing to get finer results.

6 Conclusions

In our research, we aim at understanding the major research topics outlined in six years of conferences dedicated to OSS. We have clustered articles by their cross-citations and inspected each single cluster to search for major research themes, evolution of research topics, and major initiators or synthesizers of research in OSS.

We have also introduced a social network underlying the research production. This network assumes that authors are connected by their common research interest revealed by citations as a community of practice. This network is different from a social network defined by co-authorship. In this case, links may be transversal to research topics and are generated by existing collaborations. Differently, our network may reveal hidden and potential collaborations among researchers.

Finally, we have used a semi-automated approach for systematic mapping studies that automatically creates clusters from cross-citations and manually inspect the clustered papers. Our research has identified cornerstone papers and links among them revealing unexpected and undeclared connections among authors as well as lack of connection among potentially connected topics. Large clusters are mostly initiated and sustained by empirical papers. Since 2008, synthesizers of research have introduced frameworks and platforms to perform OSS research paving the way for future work. The analysis of non-cited papers indicates that significant research has not been exploited, yet. Therefore, we recommend the OSS community to exploit further the potential provided by the OSS conference series while maintaining the interest in its major research streams.

Acknowledgements. We would like to thank Muhammad Ali Babar and Klaas Jan Stol and the reviewers for their valuable comments.

References

1. Ågerfalk, P.J., Deverell, A., Fitzgerald, B., Morgan, L.: Assessing the Role of Open Source Software in the European Secondary Software Sector: A Voice from Industry. In: Proceedings of the 1st International Conference on Open Source Systems (OSS 2005), Genoa, Italy, pp. 82–87 (2005)
2. Aksulu, A., Wade, M.R.: A Comprehensive Review and Synthesis of Open Source Research. Special Issue, Journal of Association for Information Systems 11(11), 576–656 (2010)
3. Anel, J.A.: The Importance of Reviewing the Code. Communication of the ACM, 40–41 (May 2011)
4. Ayala, C., Hauge, Ø., Conradi, R., Franch, X., Li, J., Velle, K.S.: Challenges of the Open Source Component Marketplace in the Industry. In: Boldyreff, C., Crowston, K., Lundell, B., Wasserman, A.I. (eds.) OSS 2009. IFIP AICT, vol. 299, pp. 213–224. Springer, Heidelberg (2009)

5. Biolchini, J., Mian, P. G., Natali, A. C. C., Travassos, G. H.: Systematic Review in Software Engineering, University of Rio de Janeiro:TR-ES 679/05 (2005)
6. Brereton, P., Kitchenham, B., Budgen, D., Turner, M., Khalil, M.: Lessons from applying the systematic literature review process within the software engineering domain. Journal of Systems and Software 80, 571–583 (2007)
7. Budgen, D., Charters, S., Turner, M., Brereton, P., Kitchenham, B., Linkman, S.: Investigating the applicability of the evidence-based paradigm to software engineering. In: Proceedings of the 2006 International Workshop on Workshop on Interdisciplinary Software Engineering Research, Shanghai, China, May 20 (2006)
8. Conklin, M.: Beyond Low-Hanging Fruit: Seeking the Next Generation in FLOSS Data Mining. In: Proceedings of the 2nd International Conference on Open Source Systems (OSS 2006), Como, Italy, pp. 47–56 (2006)
9. Del Bianco, V., Lavazza, L., Morasca, S., Taibi, D.: Quality of Open Source Software: The QualiPSo Trustworthiness Model. In: Boldyreff, C., Crowston, K., Lundell, B., Wasserman, A.I. (eds.) OSS 2009. IFIP AICT, vol. 299, pp. 199–212. Springer, Heidelberg (2009)
10. Dybå, T., Dingsøyr, T.: Applying Systematic Reviews to Diverse Study Types: An Experience Report. In: Proceedings of the Proceedings of the First International Symposium on Empirical Software Engineering and Measurement (2007)
11. Dybå, T., Kitchenham, B., Jorgensen, M.: Evidence-Based Software Engineering for Practitioners. IEEE Software 22, 58–65 (2005)
12. Feller, J., Fitzgerald, B.: Understanding Open Source Development. Addison-Wesley, Reading (2001)
13. Feller, J., Fitzgerald, B., Hissam, A.S., Lakhani, K.R.: Perspectives on Free and open Source Software. MIT Press, Cambridge (2007)
14. Gao, Y., Madey, R.G.: Network Analysis of the SourceForge.net Community. In: Proceedings of the 3rd International Conference on Open Source Systems (OSS 2007), Limerick, Ireland, pp. 187–200 (2007)
15. Gasser, L., Scacchi, W.: Towards a Global Research Infrastructure for Multidisciplinary Study of Free/Open Source Software Development. In: Proceedings of the 4th International Conference on Open Source Systems (OSS 2008), Milano, Italy, pp. 143–158 (2008)
16. Hauge, Ø., Ayala, C.P., Conradi, R.: Adoption of open source software in software-intensive organizations - A systematic literature review. Information & Software Technology 52(11), 1133–1154 (2010)
17. Hauge, Ø., Cruzes, D., Conradi, R., Sandanger Velle, K., Skarpenes, T.A.: Risks and Risk Mitigation in Open Source Software Adoption: Bridging the Gap between Literature and Practice. In: Ågerfalk, P., Boldyreff, C., González-Barahona, J.M., Madey, G.R., Noll, J. (eds.) OSS 2010. IFIP AICT, vol. 319, pp. 105–118. Springer, Heidelberg (2010)
18. Hauge, Ø., Ziemer, S.: Providing Commercial Open Source Software: Lessons Learned. In: Boldyreff, C., Crowston, K., Lundell, B., Wasserman, A.I. (eds.) OSS 2009. IFIP AICT, vol. 299, pp. 70–82. Springer, Heidelberg (2009)
19. Hauge, Ø., Sørensen, C., Conradi, R.: Adoption of Open Source in the Software Industry. In: Proceedings of the 4th International Conference on Open Source Systems (OSS 2008), Milano, Italy, pp. 211–221 (2008)
20. Hauge, Ø., Sørensen, C., Røsdal, A.: Surveying Industrial Roles in Open Source Software Development. In: Proceedings of the 3rd International Conference on Open Source Systems (OSS 2007), Limerick, Ireland, pp. 259–264 (2007)

21. Howison, J., Inoue, K., Crowston, K.: Social Dynamics of Free and Open Source Team Communication. In: Proceedings of the 2nd International Conference on Open Source Systems (OSS 2006), Como, Italy, pp. 319–330 (2006)
22. Howison, J., Conklin, M., Crowston, K.: OSSmole: A collaborative repository for FLOSS research data and analyses. In: Proceedings of the 1st International Conference on Open Source Systems (OSS 2005), Genoa, Italy, pp. 54–60 (2005)
23. Kitchenham, B.: What's up with software metrics? - A preliminary mapping study. Journal of Systems and Software 83(1), 37–51 (2010)
24. Kitchenham, B., Brereton, O.P., Budgen, D., Turner, M., Bailey, J., Linkman, S.: Systematic literature reviews in software engineering – A systematic literature review. Information and Software Technology 51, 7–15 (2009)
25. Kitchenham, B., Charters, S.: Guidelines for Performing Systematic Literature Reviews in Software Engineering, Keele University, UK EBSE-2007-1 (2007)
26. Kitchenham, B.: Procedures for Performing Systematic Reviews, Keele University Technical Report TR/SE-0401 (2004)
27. Koch, S.: Evolution of Open Source Software Systems – A Large-Scale Investigation. In: Proceedings of the 1st International Conference on Open Source Systems (OSS 2005), Genoa, Italy, pp. 148–153 (2005)
28. Kyo, C.K.: FODA: Twenty years of Perspectives on feature Models. In: Keynote at 13th International Product Line Conference, SPLC (2009)
29. Petersen, K., Feldt, R., Mujtaba, S., Mattsson, M.: Systematic mapping studies in software engineering. In: Proceedings of the 12th International Conference on Evaluation and Assessment in Software Engineering, pp. 71–80 (2008)
30. Robles, G., Gonzales Barahona, J.M., Michlmayr, M.: Evolution of Volunteer Participation in Libre Software Projects: Evidence from Debian. In: Proceedings of the 1st International Conference on Open Source Systems (OSS 2005), Genoa, Italy, pp. 100–107 (2005)
31. Rossi, B., Scotto, M., Sillitti, A., Succi, G.: Criteria for the non invasive transition to OpenOffice. In: Proceedings of the 1st International Conference on Open Source Systems (OSS 2005), Genoa, Italy, pp. 250–253 (2005)
32. Simmons, G.L., Dillon, T.S.: Towards an Ontology for Open Source Software Development. In: Proceedings of the 2nd International Conference on Open Source Systems (OSS 2006), Como, Italy, pp. 65–75 (2006)
33. Sirkkala, P., Aaltonen, T., Hammouda, I.: Opening Industrial Software: Planting an Onion. In: Boldyreff, C., Crowston, K., Lundell, B., Wasserman, A.I. (eds.) OSS 2009. IFIP AICT, vol. 299, pp. 57–69. Springer, Heidelberg (2009)
34. Staples, M., Niazi, M.: Experiences using systematic review guidelines. Journal of Systems and Software 80, 1425–1437 (2007)
35. Stol, K., Ali Babar, M.: A Comparison Framework for Open Source Software Evaluation Methods. In: Ågerfalk, P., Boldyreff, C., González-Barahona, J.M., Madey, G.R., Noll, J. (eds.) OSS 2010. IFIP AICT, vol. 319, pp. 389–394. Springer, Heidelberg (2010)
36. Stol, K.-J., Ali Babar, M.: Reporting empirical research in open source software: The state of practice. In: Boldyreff, C., Crowston, K., Lundell, B., Wasserman, A.I. (eds.) OSS 2009. IFIP AICT, vol. 299, pp. 156–169. Springer, Heidelberg (2009)
37. Stol, K.J., Ali Babar, M., Russo, B., Fitzgerald, B.: The use of empirical methods in Open Source Software research: Facts, trends and future directions. In: ICSE Workshop on Emerging Trends in Free/Libre/Open Source Software Research and Development, pp. 19–24 (2009)

38. Taibi, D., Lavazza, L., Morasca, S.: OpenBQR: a framework for the assessment of OSS. In: Proceedings of the 3rd International Conference on Open Source Systems (OSS 2007), Limerick, Ireland, pp. 173–186 (2007)
39. Ven, K., Verelst, J.: The Organizational Adoption of Open Source Server Software by Belgian Organizations. In: Proceedings of the 2nd International Conference on Open Source Systems (OSS 2006), Como, Italy, pp. 111–122 (2006)
40. Zennier, C., Melnik, G., Maurer, F.: On the success of empirical studies in the international conference on software engineering. In: Proceedings of the 28th International Conference on Software Engineering (ICSE 2006), Shanghai, China, pp. 341–350 (2006)