Investigating probabilistic sampling approaches for large-scale surveys in software engineering

Rafael Maiani de Mello*, Pedro Corrêa da Silva and Guilherme Horta Travassos

Abstract

Background: Establishing representative samples for Software Engineering surveys is still considered a challenge. Specialized literature often presents limitations on interpreting surveys’ results, mainly due to the use of sampling frames established by convenience and non-probabilistic criteria for sampling from them. In this sense, we argue that a strategy to support the systematic establishment of sampling frames from an adequate source of sampling can contribute to improve this scenario.

Method: A conceptual framework for supporting large scale sampling in Software Engineering surveys has been organized after performing a set of experiences on designing such strategies and gathering evidence regarding their benefits. The use of this conceptual framework based on a sampling strategy developed for supporting the replication of a survey on characteristics of agility and agile practices in software processes is depicted in this paper.

Result: A professional social network (LinkedIn) was established as the source of sampling and its groups of interest as the units for searching members to be recruited. It allowed to deal with a sampling frame composed by more than 110,000 members (prospective subjects) distributed over 19 groups of interest. Then, through the similarity levels observed among these groups, eight strata were organized and 7745 members were invited, from which 291 have confirmed participation and answered the questionnaire.

Conclusion: The heterogeneity and number of participants in this replication contributed to improve the strength of original survey’s results. Therefore, we believe the sharing of this experience, the instruments and plan can be helpful for those researchers and practitioners interested on executing large scale surveys in Software Engineering.

Keywords: Population; Sampling frame; Experimental software engineering; Hierarchical clustering analysis; Stratified sampling; Survey; Multivariate analysis; Graph theory; Strongly connected components

1 Introduction

Primary studies in Software Engineering (SE) are often conducted over samples established by convenience (Pickard et al. 1998; Sjøberg et al. 2005; Dybå et al. 2007). This scenario is especially critical for large scale surveys (Kasunic 2005), in which considerable efforts on participants recruitment and data gathering are applied (Conradi et al. 2005) but the generalization of results are limited, even when other survey’s features are clearly described and repeated in their trials. One challenge on establishing
representative samples in SE surveys includes identifying relevant and available sources from which sampling frames can be established. Therefore, SE researchers are tempted to explore alternative sources typically available on the Web for enlarging samples’ size, such as social networks (de Mello and Travassos 2013; Joorabchi et al. 2013). However, the ad hoc use of such Web technologies per se is not sufficient to evolve the sampling scenario regarding SE surveys, since the size of samples is just one of the issues hampering the generalization of SE surveys’ results. Indeed, a systematic process of sampling should be established to support the identification of representative samples from adequate sampling frames. In our research context, a sample is representative whether such sample was probabilistically extracted from a sampling frame in which is expected to observe the same heterogeneity observed in the surveys’ target audience with respect to certain known characteristics. In this sense, three meanings of the same concept presented by (Kruskal and Mosteller 1979) drive our definition of “representative sample”: specific sampling method (probabilistic sampling), coverage of the populations’ heterogeneity and representative as typical (with respect to certain known characteristics of the population). However, the two last meanings are expected to be supported, taking into account the typical limitations on characterizing the whole target audience and identifying adequate sources for sampling on SE surveys.

The technical literature regarding surveys and the specific challenges observed in SE research for characterizing population and sampling supported us in the development of a first recruitment plan concerned with the replication of a survey on requirements effort estimation (Vaz 2013; de Mello and Travassos 2013). This recruitment plan allowed the application of systematic search over a professional social network (LinkedIn) to identify a more representative sample than those previously established by convenience in the original study. After comparing both original and replicated trials’ samples (i.e., the sets of individuals that effectively answered the questionnaire), we observed that although both samples presented similar confidence levels, the replicated trial sample was significantly more heterogeneous. Therefore, this first replication through a recruitment plan supported the compilation of lessons learned and directed our investigation efforts towards the development of new and evolved recruitment plans for SE surveys.

Then, a new recruitment plan was developed and used to support the replication of another survey for gathering the opinion of SE practitioners regarding characteristics of agility and agile practices in Software Processes. This survey had also been previously executed twice with samples consisting of all authors of the papers identified through a systematic literature review undertook to organize an initial set of characteristics and agile practices (Abrantes and Travassos 2013). Such original trials allowed to receive the opinion of 25 participants (from 158 invitations), from which only seven of them declared “High” or “Very High” experience on applying agile technologies in software projects. Otherwise, the replication using the recruitment plan presented in (de Mello et al. 2014a) allowed the identification of 7745 distinct members from 19 LinkedIn (www.linkedin.com) groups of interest. As a result, the survey was answered by 291 subjects from 149 countries distributed across all continents and organized into five strata based on the reported participant’s experiences and their SE main skills (de Mello et al. 2014b). Within this heterogeneous sample, we identified relevant evidence regarding the opinion of distinct groups of SE professionals regarding the survey context (de
Mello et al. 2014c). The findings and lessons acquired in the replicated studies supported the organization of a conceptual framework for supporting large scale sampling in SE surveys as initially summarized in (de Mello et al. 2014d).

To present more detailed information and further explanations regarding the conceptual framework to support large sampling in SE surveys, this paper extends our previous ESELAW 2014 work (de Mello et al. 2014a). In this extended version, some background regarding SE surveys and discussions about the use of sources of sampling available in the Web are included to ground all the used concepts and facilitate the overall understanding and application of the proposed sampling approach. Besides, it details the recruitment plan and presents an actual instance of the conceptual framework, which can contribute to the use and replication of the sampling approach procedures by SE researchers and practitioners interested on performing surveys in SE.

Besides this introduction, this paper presents the following structure: Section 2 presents the background and literature review, highlighting guidelines available for SE surveys, discussing recent surveys and the use of alternative sources of sampling available on the Web. Section 3 presents the research design and the methodology, introducing the conceptual framework. Section 4 (method) describes the recruitment plan. In Section 5, the results from the plan execution are presented. Section 6 discusses how these results contributed to perform a better sample, delivering relevant contributions to the survey context.

2 Background

Questionnaire-based survey consists in a research method in which participants answer questions or respond to statements that were developed in advance. When properly conducted, this type of survey allows the researchers to generalize beliefs and opinions from a relevant population from the target audience by studying a subset (sample) of them. Kasunic (2005) presents the following steps for the questionnaire based survey process:

1. Identify the research objectives
2. Identify and characterize target audience
3. Design the sampling plan
4. Design and write questionnaire
5. Pilot test questionnaire
6. Distribute the questionnaire
7. Analyze results and write report

The second and third steps are the focus of our research. Step 2 is concerned with analyzing the survey objectives for extracting the target audience, driving the establishment of the survey's population and its sampling frame (Step 3) in which statistical concepts for sampling can be applied. Therefore, most of the discussions in this paper are concentrated on these two main concepts: population and sample. To support the discussions regarding our proposal, next subsections exemplifies how surveys have been conducted in SE and in which extent there are guidelines available focusing on
population and sampling issues. Studies using alternative sampling sources such as those available in the Web are also discussed.

2.1 Survey guidelines for SE studies
In a series of papers, Kitchenham and Pfleeger (2001) introduced the principles of survey research for SE researchers, mainly covering steps 3, 4, 6 and 7 of Kasunic’s proposal (2005). In the context of characterizing the population and designing the sampling plan, some basic statistical concepts are described, with few discussions regarding SE population issues being provided (Kitchenham and Pfleeger Kitchenham and Pfleeger 2002). Kitchenham and Pfleeger (2008) also investigated the design of four SE surveys conducted from 1998–2000 from which authors conclude that only one of them was supported by a representative sample from a clearly established population.

Based on their practical experiences on conducting SE surveys, Ciolkowski et al. (2003) present a comprehensive work concerned with guidelines for conducting surveys in SE. However, regarding the population and sampling, they only present how the samples were designed for their studies. Alternatively, Conradi et al. (2005) reported details on how they established survey’s target audience and obtained a representative sample through an exhaustive process of gathering organization’s data from unbiased sources from three countries (Italy, Germany and Norway). The authors also describe a relevant set of attributes collected from each individual and his/her job and how they used these attributes for analyzing the survey’s results. Moreover, this survey was replicated by Ji et al. (2008), in which the authors discuss the challenges on replicating it in China. In both studies, the authors discuss the challenges and the limitations on establishing representative samples for SE surveys. However, it was not possible to observe a proposal of solution in their report.

Finally, it is important to highlight the relevant contribution of Kasunic’s technical report (2005) for conducting surveys in SE, in which a seven-stage survey process is detailed and presented through a hands-on guideline. However, the chapters devoted to population and sampling barely deal with specific issues regarding SE surveys.

2.2 Surveys in software engineering
The technical literature in SE presents many researches using surveys as an investigation strategy. Not intending to be exhaustive and just getting some recent works, these studies are concerned with many distinct topics in SE, such as software development for medical devices (Denger et al. 2007), requirements engineering (Chen et al. 2013), software processes (Guo and Seaman 2008; Rodríguez et al. 2012; Abrantes and Travassos 2013), pair programming (Rodríguez et al. 2012), defect reporting (Bettenburg et al. 2008), exploratory testing (Pfahl et al. 2014), model-based testing (Dias Neto and Travassos 2008), effort estimation (Basten and Mellis 2011; Vaz 2013), component-based software engineering (Conradi et al. 2005; Ji et al. 2008), global software development (Humayun et al. 2013) and technology transfer (Diebold and Vetrò 2014) among many others.

Depending on the research question and its target audience and considering the set of surveys previously refereed, one can observe that the unit of analysis, i.e. the basic element of analysis from each survey (Hopkins 1982), can be typically established as:
Thus, one can see that a unit of analysis can be composed by one or more units of observation, i.e. the set of individuals composing a target audience. The trend on conducting surveys having an individual as unit of analysis in SE research could be explained due to the common restricted access to large-scale sampling frames composed by units such as software projects or organizations. However, it is also important to observe that organizations and software projects can also be used as sources of prospective subjects when individuals are, in fact, the units of analysis. This practice can be observed in surveys concerned with open source projects (Bettenburg et al. 2008) and software organizations (Humayun et al. 2013). Other approach to support the identification of prospective subjects (especially when the target audience is represented by researchers) is based on the recruitment of the papers’ authors identified through secondary studies, such as systematic literature reviews (SLR) (Dias Neto and Travassos 2008; Abrantes and Travassos 2013). In such cases, census is commonly applied since authors from all retrieved papers are invited.

It is common to observe efforts invested to avoid sampling biases in SE surveys. However, such surveys rarely evaluate the representativeness of the collected samples over their respective sampling frames. It can also be observed that few works explore the samples’ heterogeneity observed through samples’ attributes in order to better interpreting the survey’s results. In this sense, some interesting examples can be observed in the works of Conradi et al. (2005); (Basten and Mellis 2011), and Pfahl et al. (2014).

2.3 Web-based sources for sampling

Web-based sources of recruitment represent contemporary alternatives to identify populations and subjects samples. For instance, Stolee and Elbaum (2013) used a crowdsourcing tool (Mechanical Turk) for recruiting participants in large scale to an experiment on Java code search. As the tasks were opened to the “crowd”, each individual was first invited to take a brief proficiency test. Having the individual be approved, the individual was classified as “able” to participate in the experiment and perform the tasks. However, crowdsourcing tools typically hide relevant information regarding their populations, making unfeasible the clear characterization of the sampling frame (de Mello et al. 2014d). Thus, although this kind of source of recruitment potentially contributes to increase the sample’s size, little guidance is available to support the generalization of results.

In order to increase the samples’ sizes without using “convenience” as criterion, Jedlitschka et al. (2007) distributed invitations to “several websites and forums” whereas Joorabchi et al. (2013) sent invitations to “Mobile Development Meet up groups, LinkedIn groups related to native mobile development” and also “shared the survey through our Twitter accounts”. Bettenburg et al. (2008) extracted subjects from defects reported
in open source projects. However, none of these works presented an explicit sampling design or the evaluation in which extent their samples were representative.

Alternatively, de Mello and Travassos (2013) and de Mello et al. (2014a) report the use of systematic recruitment plans to support the identification of populations and representative samples for SE surveys from a source of sampling such as LinkedIn. In these cases, groups of interest were retrieved from LinkedIn through a search algorithm and, next, random samples of members were organized. As a result, researchers were able to evaluate samples’ representativeness and to identify similar sets of members and groups, delivering more accurate findings and strengthen the surveys’ results.

2.4 Research design and methodology
The research presented in this paper aims to depict the use of a conceptual framework to support large-scale sampling in SE surveys. This work evolves the original recruitment plan (de Mello et al. 2014a) and presents the first version of a Conceptual Framework proposed by de Mello et al. (2014d) which is represented in Fig. 1. In addition to the statistical concepts of target audience, population, sampling frame and unit of observation (Thompson 2012), this framework introduces the following set of new concepts for better supporting sampling in SE surveys: source of sampling, search unit, search plan and sampling strategy. A source of sampling consists on a database (automated or not) in which valid subpopulations of the target audience can be systematically retrieved and randomly sampled. Thus, if a source of sampling can be considered valid for a specific research context, it can be concluded that sampling frames can be

![Fig. 1 Conceptual framework for large scale sampling in SE surveys (de Mello et al. 2014d)]
established from it for the same research context. To be considered valid, a source of sampling should satisfy, at least, the following essential requirements (ER):

- **ER1.** A source of sampling should not intentionally represent a segregated subset from the target audience, i.e., for a target audience “X”, it is not adequate to search for units from a source intentionally designed to compose a specific subset of “X”.
- **ER2.** A source of sampling should not present any bias on including on its database preferentially only subsets from the target audience. Unequal criteria for including search units mean unequal sampling opportunities.
- **ER3.** All source of sampling’s search units and their units of observation must be identified by a logical or numerical id.
- **ER4.** All source of sampling’s search units must be accessible. If there are hidden search units, it is not possible to contextualize the population.

There are also nine desirable requirements (DR), three concerned with the samples’ accuracy (ADR), two concerned with clearness (CDR) and four regarding sample’s completeness (CoDR). These additional criteria and examples of evaluating such sources using them can be found in (de Mello et al. 2014d).

The search unit characterizes how one or more units of observation can be retrieved from a specific source of sampling. In an ideal scenario, it is expected that both unit of observation and search unit are as much as possible the same. Search plan describes how search units will be systematically retrieved from a source of sampling and evaluated in order to compose a sampling frame. Finally, the sampling strategy describes the steps that must be followed for sampling and recruiting individuals that will take part in the study trial. Eventually, the data used for supporting the sampling design can be retrieved after collecting answers with a measurement instrument such as a characterization form, before executing the survey.

For instance, to perform a survey from which the target audience is composed by Brazilian D.Sc. students in SE, one of the relevant sources of sampling is the CNPq research group directory (http://dgp.cnpq.br/dgp/). Thus, the search unit can be defined as each research group, whereas each unit of observation can be defined as each D.Sc. student from each selected group. Then, for the search plan, in order to avoid groups out of context, a search algorithm can be performed applying the search expression “software engineering” for selecting only those research groups concerned with the target audience. As a result, it is expected to establish a sampling frame composed by all D.Sc. students retrieved from the selected research groups. Then, simple random sampling (SRS) can be defined as the sampling strategy, aiming at recruiting an amount of subjects (sample size) to support, in the worst case, a confidence level of 95 % and a confidence interval of 2.00. Thus, if all recruited subjects answer this survey, in the case of the observed proportion for a response in a survey question be 50 % (worst case), there is a probability of 95 % that this result will be repeated to the whole population having a margin of error of 2 points (48 % ~ 52 %).

### 3 Method

The replication of the survey on characteristics of agility and agile practices in software processes (Abrantes and Travassos 2013) is going to illustrate the use of a systematic
recruitment plan and its contribution to retrieve representative samples from a professional social network (LinkedIn). In the case of this survey, the target audience is composed by SE practitioners in general since a set of characteristics and practices identified as “agile” in SE may be evaluated independent from the software process adopted. Thus, all professionals working in software projects can potentially contribute with this investigation. It is important to highlight that the survey plan weights the participant’s relevance and correspondingly answers by his/her experience level. The following subsections describe the recruitment strategy designed for this survey trial.

3.1 Source of sampling, search unit and population
A professional social network (LinkedIn) has been established as source of sampling due its coverage, consisting of more than 10 million of IT workers spread in the world (November 2014). For performing the recruitment plan and the data analysis presented in this study, the use of a “Premium” account was necessary. This account type allows the LinkedIn’s users to perform more accurate analysis regarding the distribution of members between groups of interest. Since LinkedIn allows performing a comprehensive group of interests searching, “group of interest” will be the search unit. From each identified group, it will be extracted the following attributes: Group Name, Group Description, Group Size (amount of members) and Group Official Language. These attributes will be used to verify whether each group of interest can be included in the sampling frame, which is expected to be composed by all groups of interest concerned with agility in software process. Thus, the population from this survey trial will be composed by all members from these selected groups.

3.2 Unit of observation and unit of analysis
In this survey, the unit of observation and unit of analysis are the same entity (individual) and each distinct member from each group are potentially considered a valid unit to be sampled. The following attributes should be collected from each one:

- Attributes collected through the source of sampling: Member ID, Name, Country and Status of Membership in each group of interest in the sampling frame. One can see that individuals’ profiles in LinkedIn present other attributes that can be used in our investigation, such as academic degree, professional experience and top skills. However, these data is commonly not accessible when the individual profile is not directly connected with user account. Also, there is no control regarding whether these attributes are updated.
- Attributes collected through the survey’s questionnaire (measurement instrument): Country, Main Skills in SE, SE Experience Level, Agility Experience Level and Academic Degree.

3.3 Search plan
The search question for establishing the sampling frame is: “Which are the groups from LinkedIn related to agility in software processes?” Thus, based on the search string from the SLR and its results (Abrantes and Travassos 2013), the following search expressions were established:
“agile”, “agility”, “test-driven development”, “continuous integration”, “pair programming”, “planning game”, “on site customer”, “collective code Ownership”, “collective ownership”, “small releases”, “short release”, “developing by feature”, “metaphor”, “refactoring”, “Sustainable Pace”, “simple design”, “coding standards”, “whole team”, “project visibility”, “daily meetings”, “open workspace”, “product backlog”

The steps for applying the searches concerned with the presented search expressions were defined in the following search algorithm:

For each keyword, do:

Submit a search expression (between quotes) followed by the term “software” in the option “Group Search”;
Identify all groups of interest returned, recovering the following data: name, description, group rules and number of members.

Then, aiming at restricting the selection of groups of interest to those discussing agile practices and characteristics in the global context, it will be excluded any group of interest that:

- explicitly prohibits the execution of studies;
- explicitly restricts the individual messaging between its members (a default feature provided by LinkedIn);
- is explicitly directed to a city, region or country, since our target audience are not geographically restricted;
- is focused on promoting specific organizations, or provided by them, neither to disseminate specific events;
- has its description out of the scope of Software Engineering;
- has a vague description;
- has a single member;
- is driven to headhunting and job offering;
- represents LinkedIn' subgroups, since the sampling frame must be composed by groups of interest, and;
- has a non-English language as default, since English language is default in international forums.

3.4 Sampling strategy

Based on groups of interest’s similarities (groups’ members overlapping), the feasibility of performing the following sampling designs will be analyzed, in this order:

1. Clustered sampling: this sampling design can be applied when homogeneous groups (clusters) composed by distinct units can be identified in a population. As a consequence, due to this similarity (identified as function of a set of units’ attributes), only a subset from these clusters can be randomly sampled without significant loss of confidence, also reducing efforts on recruitment and data collection (Thompson, 2012). Thus, clustered sampling is commonly applied in
large scale surveys in which researchers have operational restrictions for recruiting and collecting data (Bennett et al. 1991; Roberts et al. 2004; Eldridge et al. 2006).

2. Stratified sampling: it is considered the best probabilistic sampling design, taking into account the heterogeneity from each population and distributing its units into distinct subpopulations (Thompson, 2012). Then, if these subpopulations (strata) are well established, it can be considered that the population from each stratum is homogeneous for the study context and SRS may be performed in each one, allowing the observation of more specific and reliable results than a single SRS from the whole population. In stratified sampling, it is expected mutual exclusion between the units from each stratum.

Independent from the sampling design used, it will be randomly extracted a sample size from each group of interest in order to cover, in the worst case, a confidence interval of 3.50 (margin of error) for a confidence level of 99 %. However, due to the low participation rate (3.7 %) observed in our first study (de Mello and Travassos 2013), it is expected to support the results in providing significantly lower confidence.

3.5 Instrumentation

For each selected and included group of interest from the source of sampling, it will be analyzed the coverage of this group over the entire population, in order to establish a sampling frame. The coverage from each group G will be calculated as the ratio between the number of members from G and the amount of distinct members from all selected groups. Then, for each group included in the sampling frame, the researcher with a LinkedIn “Premium” account will send a subscription request, since only using this type of account it is possible to identify groups members’ overlapping, supporting the definition of the net population size. If accepted, the group of interest will be preserved in the sampling frame. Then a default recruitment message will be individually sent using exclusively the message service provided by LinkedIn, the source of sampling.

3.6 Automated support

In order to make feasible and mitigate the operational risks on manually selecting and sending a large amount of invitations from LinkedIn, it was provided automated support to data extraction from the members’ profile and messages sending activities. Figure 2 shows the components composing these activities in the same sequence they are used.

3.7 Data extraction

To extract the required attributes from each member (section 4.2), it was used the Firefox add-on named iMacros (addons.mozilla.org). This plugin makes possible to select values from HTML tags and save them as CSV (Comma Separated Values) file. Figure 3 demonstrate how iMacros detects the desired information (in this case the member’s name).

For instance, a HTML tag < a > and its class value (title, i.e., member name) are provided. Then, to extract each member’s name, it must be informed this tag to iMacros (Fig. 3, bottom right) which can be supported through the Firefox’ Inspector tool (Fig. 3,
A similar procedure is used to extract all other relevant attributes from each member to compose the *iMacros* script. The following code exemplifies this script. First line contains a tag value (title) that must be extracted from each member and the second line shows the file in which data will be saved:

```
TAG POS = 1 TYPE = a ATTR = class:title EXTRACT = TXT
SAVEAS TYPE = EXTRACT FOLDER = "C:\" FILE = DataExtracted.txt
```

Thus, after the definition of *iMacros* script, the data extraction activity can be started (Fig. 1). First, (A) the script is executed, looping across each member listed on each search results page from *LinkedIn*, collecting all attributes that must be selected from the source of sampling (Section 4.2). Then, a CSV file (B) stores all the raw data retrieved, such as the following:

![Fig. 2 Components provided to automate data extraction and recruitment activities](image)

![Fig. 3 Example of extracting relevant attributes using Inspector tool and iMacros](image)
As it can be seen, the CSV file contains many undesired data and text artifacts that should be removed. Thus, we implemented a Ruby script (C) to remove any unnecessary data from each member, generating a final CSV file (D) composed by lines such as the following:

Rafael Maiani de Mello, 111111, CIO at Aliansce Shopping Center S.A., 1, Brazil

Finally, this CSV file is used to feed a PostgreSQL database (E).

3.8 Sending the messages

After performing the data extraction, the sending of messages can be performed, as showed in Fig. 2. Thus, from the database (E), it will be possible to randomly select the members to be recruited. The result of this recruitment will be a new CSV file (F) containing the following attributes from the selected members: user id, group id and member name. Then, a second iMacros Script takes in action using the information available in (F), generating an individual message containing a personalized parameterized URL of the survey for each member. These messages are filled (one-by-one) in LinkedIn’s message form and sent for each member (G) based on the following template:

“Dear <Member Name>
I’m Rafael de Mello (http://www.cos.ufrj.br/~rmaiani), Phd. Student at COPPE/UFRJ, Brazil. I'm member of the Experimental Software Engineering Group (http://ese.cos.ufrj.br/en), supervised by Prof. Guilherme Horta Travassos (http://www.cos.ufrj.br/~ght).
Since 2009, our research group is conducting researches concerned with agility in software processes. As part of them, Dr. Jose Abrantes planned and executed a survey (two trials), which has identified a set of agile characteristics and practices applicable in software processes. At this time, we are re-applying this survey (third trial) aiming at to reach a large-scale population of researchers/practitioners interested on this topic. Based on your area of interest, we kindly invite you to take part in the following survey:
http://lens-ese.cos.ufrj.br/SurveyAgile/index.php?userID=<UserID>&groupID=<GroupID>
Your opinion is essential to strength our findings. Please, help us accordingly your possibilities by answering this survey until December 13th.
As soon as we conclude data analysis, we will share the results with all participants and the software engineering community.
Thanks in advance,
Rafael de Mello”

4 Result

In September 2013, 289 distinct groups of interest were retrieved applying the search algorithm and 227 were excluded due to one or more exclusion criterion (Table 1). It is
important to highlight that all 289 groups of interest neither prohibit the execution of studies nor restrict the sending of individual messages among their members, a type of resource provided by LinkedIn by default. In fact, each member from a group of interest can chose, at any time, to receive or not messages from members of the same groups of interest.

After applying the exclusion criteria, 62 groups of interest were selected. Then, it was observed that more than 92 % of the total gross population (including the repetition of members on each group of interest) is presented over the third quartile from the distribution of groups’ amount of members (Table 2, groups from “A” to “P”). Then, only these 16 groups were selected to compose the sampling frame. However, since the researcher’s profile was not accepted in the groups G and O, these groups of interest were substituted by the groups Q, R, S, T and V in which the same researcher was immediately accepted. After this substitution, the coverage of 90 % from all gross population was preserved. Table 3 presents the final sampling frame.

One can see by the group of interest’s names (Table 3) that some of them are not explicitly related with the agile context. In fact, although the source of LinkedIn’ group search algorithm is not available, we observed that it searches the occurrence of the search expressions not only in the groups of interest’s name and description, but also in the groups of interest’s discussions information. However, detailed heuristics regarding LinkedIn filtering methods are not available.

In November 2013 the sampling frame totalized 264,540 gross members, distributed over 149 countries. Using the filtering resources from LinkedIn, 202,643 distinct

---

**Table 1** Incidence of each exclusion criteria

| Exclusion criteria                        | #  | % from Total |
|-------------------------------------------|----|--------------|
| Local Groups                              | 97 | 42.73 %      |
| Organizations, publicity and events       | 66 | 29.07 %      |
| Out of scope                              | 33 | 14.54 %      |
| Vague description                         | 25 | 11.01 %      |
| Single member groups                      | 18 | 7.93 %       |
| Headhunting and job offering groups       | 8  | 3.52 %       |
| LinkedIn subgroups                        | 3  | 1.32 %       |
| Non-English                               | 1  | 0.44 %       |
| Total of Excluded Groups                  | 227| 78.55 %      |

---

**Table 2** Distribution of the Selected Groups of Interest over the median (218 members)

| Group | Size  | Group | Size  | Group | Size  | Group | Size  |
|-------|-------|-------|-------|-------|-------|-------|-------|
| A     | 47,678| I     | 8,225 | Q     | 5,464 | Y     | 421   |
| B     | 39,770| J     | 7,666 | R     | 4,043 | Z     | 399   |
| C     | 37,116| K     | 7,324 | S     | 3,780 | AA    | 373   |
| D     | 26,783| L     | 6,931 | T     | 2,359 | AB    | 258   |
| E     | 21,009| M     | 5,961 | U     | 1,747 | AC    | 258   |
| F     | 20,967| N     | 5,710 | V     | 913   | AD    | 237   |
| G     | 13,958| O     | 5,690 | W     | 747   | AE    | 235   |
| H     | 9,940 | P     | 5,653 | X     | 731   |       |       |
members were identified, i.e., an overlapping rate of 30.54 %, having a great geographical concentration of members from USA and Europe. The high distribution in Asia was mainly due to India. Table 4 presents the distribution of distinct members over great regions of the world, considering the 60 more represented countries (98.23 % from total of members) in LinkedIn.

The following subsections presents the grouping analysis performed (clustered sampling and stratified sampling) over the sampling frame, as planned in the subsection 4.7.

### 4.1 Clustered sampling

Clustered sampling is an alternative sampling approach commonly used to reduce the recruitment effort and gather data from subjects, without reducing the whole population representativeness. This approach is frequently applied in medicine (Bennett et al.

| Group | Name                                                                 |
|-------|----------------------------------------------------------------------|
| A     | Agile and Lean Software                                             |
| B     | Test automation                                                      |
| C     | Scrum Practitioners                                                  |
| D     | Agile                                                                |
| E     | Agile Project Management Group                                       |
| F     | Bug Free: Discussions in Software Testing                            |
| H     | Configuration and Release Management                                |
| I     | Configuration management                                             |
| J     | Agile Testing                                                        |
| K     | Software Architect Network                                           |
| L     | Lean Agile Software Development Community                            |
| M     | Scrum Practitioners, Scrum Masters                                  |
| N     | SCM (Software Configuration Management) Professional Network         |
| P     | Software Engineers in Test                                           |
| Q     | The Agile Project Management Hub                                    |
| R     | Agilists                                                             |
| S     | eXtreme Programming (XP)                                             |
| T     | Software Refactoring                                                 |
| U     | Test Driven Developers                                               |

| Table 3 Groups included in the sampling frame |

| Great Region          | % of Members |
|-----------------------|--------------|
| USA and Canada        | 39.77 %      |
| Europe                | 29.55 %      |
| Asia                  | 20.12 %      |
| Latin America         | 4.09 %       |
| Oceania               | 3.49 %       |
| Africa                | 1.20 %       |

| Table 4 Distribution of distinct members subscribed in the sampling frame, considering its 60 most represented countries |

| Great Region          | % of Members |
|-----------------------|--------------|
| USA and Canada        | 39.77 %      |
| Europe                | 29.55 %      |
| Asia                  | 20.12 %      |
| Latin America         | 4.09 %       |
| Oceania               | 3.49 %       |
| Africa                | 1.20 %       |
1991; Roberts et al. 2004; Eldridge et al. 2006), especially when it is necessary to access risk areas to perform observations. If groups composed by similar areas are mapped, just a random sample from these groups must be accessed.

In the case of the presented study, due to the limitations from the source of sampling to recover detailed and updated profiles from its members, the similarity levels were calculated as the overlapping rate between the 19 groups of interest, i.e., if a group of interest “X” has an overlapping distribution with other 18 groups of interest, similar with the distribution of a group of interest “Y”, we can amalgamate them into a single cluster. Thus, the overlapping rate (S) from each group of interest “i” over each group of interest “j” was calculated according to the following formula (1):

\[
Sim_{ij} = \left( \frac{PG_{i} + PG_{j} - PN_{ij}}{PG_{i}} \right)
\]

where PG is the total number of members in a single group of interest (gross size) and PN is the total number of distinct members between two groups of interest (net size). Figure 4 shows an extract from the resultant matrix of overlapping (groups from “A” to “I”). Based on this matrix, it was performed a hierarchical clustering analysis, a kind of multivariate analysis supported by the resource “cluster observations” available on MiniTab (www.minitab.com). In this analysis, the “Average” linkage and Pearson’s correlation were applied to calculate the distances between clusters. Thus, the dendogram showed on the Fig. 4 represents the points in which clusters were performed until performing a single cluster.

Considering the high level of similarity between the variables necessary to discard one or more sets, a minimum similarity level of 75 % was established to perform a cluster. However, as could be observed on Fig. 5, all clusters were performed below this minimum level. In fact, the higher level of similarity found was between the groups “H” and “I” (50.35 %). Thus, these observations suggest that clustered sampling is unfeasible for this sampling frame.

### 4.2 Stratified sampling

As no methods for stratification in similar contexts (social networks and software engineering) was found in the technical literature, its feasibility was evaluated through two distinct approaches: the hierarchical clustering analysis presented in the previous section (5.1) and the analysis through digraphs, aiming at identifying the most strongly connected groups.
4.3 Hierarchical clustering analysis

Although stratified sampling is commonly used to organize the units of observation naturally distributed over non-overlapped strata, such as geographical regions (Thompson 2012), the expected overlapping of subjects in the context of social networks groups can suggest some affinity among them, supporting the re-organization of the sampling frame into a smaller set of groups. The clustering trial presented in the section 5.1 suggests an initial amalgamation behavior between groups of interest related with similar themes, such as: agile practices and methods (A and C; R and M; J and S); software testing (B and F) and configuration management (H and I). Thus, considering the conceptual differences between clustering and stratification approaches, a reduced minimum level of similarity (25 %) for amalgamating groups of interest into one stratum was established. As a result, seven strata were identified. Table 5 presents each stratum and its composition.

One can see that stratification in this context allows the possibility of each member being included in more than one stratum, an undesired behavior in typical sampling frames used for performing stratified sampling. However, we argue that high overlapping rates in two or more groups from a social network may suggest similarity between them (de Mello et al. 2014a). Then, to avoid bias in the sampling process, each member can be sampled only once and his/her participation will be computed only to the stratum from which he/she was recruited.

4.4 Digraph analysis

As an alternative for the analysis presented in the previous subsection, a digraph analysis was performed, in order to identify the strongest connections between the 19 groups of interest (two-by-two). For this, each group of interest was represented as a node having their relationships with other groups of interest weighted by the previously calculated overlapping rates. Thus, similarity in this case has been observed by considering the connectivity between the pairs of groups of interest. However, considering that all the 19 groups of interest are connected with each other in both directions, we filtered only the greatest overlapping rates from all 342 identified connections.
Thus, considering the low rates observed in all quartiles from the distribution of overlapping rates (0.56 %, 1.98 % and 6.14 %), only the 83 overlapping rates fitting over the third quartile (6.14 %) was represented in the digraph.

Then, a Strongly Connected Component (SCC) algorithm was performed, using GraphViz tool (www.graphviz.org). As a result, the original digraph was derived into six disjointed components as shown in Fig. 6: one digraph with 11 nodes (A, C, D, E, L, R, M, J, S, T, U); a second digraph composed by three nodes (H, I and N); a third digraph with two nodes (B and F)); node K; node P; and node Q. Thus, based on this approach, it was identified the six strata presented on Table 6.

### 4.5 Comparison between analyses

Making a comparison between the analyses presented in the subsections 5.2.1 and 5.2.2, one can see that although the digraph analysis has suggested fewer aggregated groups, it presents similar results when compared with the hierarchical clustering analysis. As main difference, one can see that the digraph analysis unifies three strata referred to agility (EM1) and agile practices (EM3 and EM4) into a single stratum (ED1). At the same time, the digraph analysis divided the stratum from hierarchical clustering analysis related to software testing (EM3) into two strata (ED3 and ED4). Thus, aiming at avoiding overestimated groupings, all the identical strata in both analyses and only the different strata having fewer groups of interest was preserved. The result is the set of strata presented in Table 7.
4.6 Sampling and Recruitment

Considering the stratified sampling (Table 7), the computing of independent sample sizes for each stratum is necessary and must conform the same confidence interval and confidence level defined in the sampling strategy (subsection 4.4). Thus, based in the number of prospective subjects in each stratum and considering the source of sampling’s technical limitations on randomly recovering members from its groups of interest, it was used a confidence level of 99% and a confidence interval of 3.5 to calculate the sample size for each stratum, resulting in the number of subjects showed in Table 8 for recruitment, totaling 7745 members proportionally distributed.

However, even using a LinkedIn Premium account a user cannot retrieve more than 700 members from each group of interest. Thus, the default search into groups of interest was only sufficient to retrieve subsets of distinct members in the strata E1, E4 and E7. Then, aiming at enlarging the number of members retrieved for the other strata having less groups of interest (E2, E3, E5, E6 and E8), a search filter was added (members with at least 10 years of professional experience/ members with less than 10 years of professional experience) and allowed to potentially retrieve 1400 members from each group of interest composing such strata.
Table 6  The six strata obtained with digraph analysis

| Stratum | Name                          | Groups                                                                 |
|---------|-------------------------------|------------------------------------------------------------------------|
| ED1     | Agility and Agile Practices   | Agile and Lean Software Scrum Practitioners Agile Agile Project Management Group Lean Agile Software Development Community Agilists Scrum Practitioners, Scrum Masters Agile Testing eXtreme Programming (XP) Software Refactoring Test Driven Developers |
| ED2     | Project Management            | The Agile Project Management Hub                                      |
| ED3     | Software Testing 1            | Test automation Bug Free : Discussions in Software Testing              |
| ED4     | Software Testing 2            | Software Engineers in Test                                             |
| ED5     | Configuration Management      | Configuration and Release Management Configuration management SCM (Software Configuration Management) Professional Network |
| ED6     | Software Architecture         | Software Architect Network                                             |

Table 7  The final stratification proposed

| Stratum | Name                  | Groups                                                                 |
|---------|-----------------------|------------------------------------------------------------------------|
| E1      | Agility               | Agile and Lean Software Scrum Practitioners Agile Agile Project Management Group Lean Agile Software Development Community Agilists Scrum Practitioners, Scrum Masters |
| E2      | Project Management    | The Agile Project Management Hub                                      |
| E3      | Agile Practices 1     | Agile Testing eXtreme Programming (XP)                                 |
| E4      | Agile Practices 2     | Software Refactoring Test Driven Developers                             |
| E5      | Software Testing 1    | Test automation Bug Free : Discussions in Software Testing              |
| E6      | Software Testing 2    | Software Engineers in Test                                             |
| E7      | Conf. Management      | Configuration and Release Management Configuration management SCM (Software Configuration Management) Professional Network |
| E8      | Software Architecture | Software Architect Network                                             |
It is important to highlight that LinkedIn presents a biased behavior on searching, preferentially retrieving its members attending the search criteria having some level of “parentage” with the user account, i.e. members having direct connections with the user (first level) or having connections with members connected with the user (second level). In order to avoid the discarding of these members, since they are part of the population, we included them on the strata and, at the same time, evaluated the impact derived from this behavior for the samples’ heterogeneity. For this evaluation, we considered the ten countries with higher frequency of members over the whole sampling frame. Tables 9 and 10 show the distribution of the groups’ members among these countries in each stratum and the retrieved members’ distribution among the same countries, respectively. Based on both distributions, Table 11 shows the correlations calculated in order to observe how much the countries’ distribution from the retrieved sample is similar to the sampling frame, considering a confidence coefficient equals to 95%.

Although the mentioned bias, the levels of correlation for the strata E2, E4, E6, E7 and E8 are high. Analyzing the probably influence of the “affinity” over the “bad” correlations obtained for E1, E3 and E5, it was observed that these three strata actually have the most influence of first-connection and second-connection members, especially in the case of E1, in which 34.26% of its retrieved members have some level of “affinity”. Thus, we experimented removing this set of “relatives” from these strata. However, when recalculated their respective correlations, only E1 was significantly influenced by this removing. Thus, we decided to apply the referred change only to E1.

| Table 8  | Sample size for each stratum and the respective number of member retrieved |
|----------|-------------------------------------------------------------------------|
| Stratum  | #Distinct members | Sample size | Retrieved members |
| E1       | 114,827           | 1,031       | 2,563             |
| E2       | 5,488             | 874         | 1,401             |
| E3       | 11,633            | 955         | 1,444             |
| E4       | 3,864             | 820         | 883               |
| E5       | 56,400            | 1,021       | 1,358             |
| E6       | 5,791             | 882         | 1,391             |
| E7       | 17,234            | 981         | 1,016             |
| E8       | 7,335             | 911         | 1,438             |

| Table 9  | Sampling frame members by ten most represented countries |
|----------|---------------------------------------------------------|
| Stratum  | USA | IND | UK | CAN | AUS | BRA | NET | SWE | FRA | CHI |
| 1        | 37,557 | 9,337 | 11,346 | 4,355 | 3,878 | 3,228 | 3,065 | 2,509 | 2,305 | 971 |
| 2        | 1547 | 535 | 471 | 274 | 238 | 199 | 131 | 115 | 117 | 90 |
| 3        | 2,236 | 1,811 | 2,253 | 351 | 488 | 486 | 374 | 244 | 322 | 74 |
| 4        | 939 | 328 | 406 | 177 | 89 | 155 | 108 | 77 | 77 | 38 |
| 5        | 18,692 | 14,942 | 3,743 | 2,115 | 1,192 | 606 | 814 | 644 | 456 | 941 |
| 6        | 3,132 | 1,069 | 235 | 132 | 43 | 25 | 20 | 57 | 16 | 245 |
| 7        | 7,723 | 2,902 | 1,501 | 683 | 407 | 155 | 316 | 353 | 250 | 202 |
| 8        | 2,371 | 1,129 | 353 | 323 | 120 | 160 | 280 | 67 | 128 | 79 |
4.8 Recruitment

At least 10% more members were retrieved than the sample size needed for each stratum as shown in Table 12. Based on the final list of retrieved members from each stratum, members were randomly ordered using a tool from Random.org (www.random.org), selecting the “n” first distinct members from each stratum, aiming at obtaining the expected sample size, as described in Table 12. Due to the possibility of finding a same member in distinct strata, the selection was performed from the stratum having less retrieved members (E7) to the stratum having more retrieved members (E1). After 15 days of the recruitment (invitation), 291 valid contributions were obtained (de Mello et al. 2014b), composing the actual sample sizes presented in Table 12. Besides, the researchers received more than 50 individual messages supporting the research. In the other hand, just two members claimed that wouldn’t like to receive any new future messages regarding this study.

Regarding the automated recruitment support, it is important to notice the need to manually restarting the message sending after each eventual error reported by iMacros on performing the sending of individual messages. However, at the end, all 7745 were successfully recruited (invited).

5 Discussion

Since stratified sampling was performed based on overlapping rates of members between groups of interest, a new grouping analysis after the survey execution was performed based on the 291 answers concerned with the subjects’ experience level (calculated from the answers to mandatory closed questions) and their informed five

| Stratum | USA | IND | UK | CAN | AUS | BRA | NET | SWE | FRA | CHI |
|---------|-----|-----|----|-----|-----|-----|-----|-----|-----|-----|
| 1       | 1,016 | 38  | 166 | 91  | 32  | 629 | 48  | 35  | 25  | 48  |
| 2       | 428   | 118 | 127 | 70  | 55  | 55  | 39  | 28  | 30  | 39  |
| 3       | 440   | 51  | 245 | 46  | 38  | 110 | 62  | 42  | 63  | 0   |
| 4       | 265   | 30  | 86  | 48  | 18  | 31  | 39  | 29  | 30  | 3   |
| 5       | 845   | 42  | 54  | 45  | 11  | 96  | 16  | 5   | 16  | 3   |
| 6       | 610   | 87  | 47  | 32  | 7   | 6   | 5   | 14  | 4   | 14  |
| 7       | 601   | 28  | 65  | 44  | 8   | 26  | 30  | 16  | 17  | 0   |
| 8       | 521   | 164 | 52  | 55  | 20  | 59  | 59  | 17  | 21  | 4   |

Table 10 Retrieved members from the ten most represented countries

Table 11 Correlation between the distributions showed in Table 9 and 10

| Stratum | Correlation | 95% (inferior) | 95% (superior) |
|---------|-------------|----------------|----------------|
| E1      | 0.8022      | 0.3487         | 0.9513         |
| E2      | 0.9957      | 0.9814         | 0.999          |
| E3      | 0.7857      | 0.309          | 0.9469         |
| E4      | 0.9595      | 0.8333         | 0.9906         |
| E5      | 0.7611      | 0.2524         | 0.9402         |
| E6      | 0.9786      | 0.9092         | 0.9951         |
| E7      | 0.943       | 0.7714         | 0.9868         |
| E8      | 0.9843      | 0.9329         | 0.9964         |
main SE skills (open question answered by 277 participants). Each one of the 1320 collected skills was coded into a SE skill's table, distributed into 88 SE skill groups. Thus, the 20 more relevant skill groups (presented in Table 13) were used to analyze skills' distributions similarities among the eight strata as reported in (de Mello et al. 2014b).

Although it was not possible to identify significant differences in the stratum’ experience levels, we identified relevant similarities between some strata based on the distribution of the mentioned 20 SE skill groups, allowing us to reorganize the eight strata into the following five amalgamated groups (de Mello et al. 2014b):

Table 12 Sample size for each stratum and the respective number of member retrieved

| Strata Name       | #Distinct Members | Recruited Sample Size | Actual Sample Size | CI for CL = 95 % |
|-------------------|-------------------|-----------------------|--------------------|------------------|
| E1 Agility        | 114,827           | 1,031                 | 57                 | 12.98            |
| E2 Project Management | 5,488          | 874                   | 40                 | 15.44            |
| E3 Agile Practices1 | 11,633          | 955                   | 56                 | 13.06            |
| E4 Agile Practices2 | 3,864           | 820                   | 35                 | 16.49            |
| E5 Software Testing 1 | 56,400         | 1,021                 | 26                 | 19.22            |
| E6 Software Testing 2 | 5,791           | 882                   | 22                 | 20.86            |
| E7 Configuration Management | 17,234    | 981                   | 23                 | 20.88            |
| E8 SW Architecture | 7,335            | 911                   | 31                 | 17.57            |

Table 13 The 20 main skill groups mapped, based on the subjects answers

| Skill group                | Skills examples                                         | Incidence (%) |
|----------------------------|---------------------------------------------------------|---------------|
| Personal Skill             | Creativity, Detailing, Learning, Planning                | 10.56 %       |
| Programming                | Algorithms, Programming Languages                        | 8.80 %        |
| SW Analysis and Design     | Object-Oriented Design, Design Patterns                  | 8.25 %        |
| Social Skill               | Communication, Leadership                               | 7.78 %        |
| SW Testing                 | Testing, Debugging                                      | 7.71 %        |
| Thinking and Reasoning     | Abstraction, Analytical Thinking                         | 6.24 %        |
| Agile Practices            | Refactoring, Test Driven Development                     | 5.05 %        |
| Agile Characteristic       | Adaptability, Being Collaborative                        | 5.00 %        |
| SW Requirements            | Req. Analysis, Requirements Elicitation                  | 4.52 %        |
| SW Quality                 | Quality, Quality Assurance                              | 3.65 %        |
| SW Architecture            | SW Architecture                                         | 3.63 %        |
| Problem Solving            | Problem Solving                                         | 3.31 %        |
| Agile Methods              | Kanban, Scrum, Extreme Programming                      | 2.71 %        |
| Business Analysis          | Business understanding, Business Analysis                | 2.66 %        |
| Project Management         | Project Management                                       | 2.21 %        |
| Technical Expertise        | Technical Knowledge                                      | 2.06 %        |
| Configuration Management   | Change Management, Release Management                    | 2.01 %        |
| Agile                      | Agile coaching, Agile thinking, Agility                  | 1.91 %        |
| SW Development Process     | SW Process Improvement, SW Development Life-Cycle        | 1.27 %        |
| SW Development             | Development, SW Development                              | 1.12 %        |
- Agilists (E1 + E2): composed by many LinkedIn groups of interest concerned with agility in SE. Its main skill groups are: Personal Skills (11.06 %), Social Skills (10.38 %) and SW Analysis and Design (9.10 %).
- Testing Professionals (E3 + E5): mainly composed by LinkedIn groups of interest devoted to Software Testing, which also represents the most relevant skill group (14.80 %);
- Programmers (E4 + E6): mainly composed by LinkedIn groups of interest devoted to agile practices, having programming as the most relevant skill group (15.76 %);
- Configuration Managers (E7): composed by three LinkedIn groups of interest devoted to configuration management (CM). Its main skills groups are CM (12.30 %), Programming (10.73 %) and Personal Skills (10.09 %).
- System Analysts (E8): composed by a single LinkedIn group of interest devoted to software architecture. Sample’s main skill groups are personal skills (15.53 %) and SW analysis and design (14.26 %).

To evaluate in which extent the survey results represent the opinion of each amalgamated group, we calculated the Confidence Interval (CI) for a confidence level of 95 % considering the worst case, as shown in Table 14. As higher the confidence interval as more imprecise are the results. Although these low participation rates, it was observed that the confidence levels presented in Table 14 are similar to the most of surveys having individuals as unit of analysis mentioned in Section 2.2. It can be explained due to the fact that such surveys typically works with small populations established by convenience which tends to increase the participation rate but not sufficiently for bringing high confidence to the results For instance, considering both first trials from the same survey, while the participation rate were significantly higher (17.95 %) its confidence interval (19.45) was only higher than the “Configuration Management” stratum (20.42).

To compare the results from the previous two trials with the results from the trial presented in this work, it was performed an aggregation of the results found in each amalgamated group weighted by its population size (de Mello et al. 2014c). The survey asked the opinion of the participants regarding the pertinence and relevance of a set of characteristics of agility and agile practices in software processes (Abrantes and Travassos 2013). In this context, it was observed the following main findings of this study as discussed in (de Mello et al. 2014c):

- The practice of metaphor and the characteristic of emergence was discarded by all amalgamated groups;

| Skill group               | Population size | Sample size | CI for CL=95 % |
|--------------------------|-----------------|-------------|----------------|
| Agilists                 | 120,315         | 97          | 9.95           |
| Testing Professionals    | 68,033          | 82          | 10.82          |
| Programmers              | 9,655           | 57          | 12.94          |
| Configuration Managers   | 17,234          | 23          | 20.42          |
| System Analysts          | 7,335           | 31          | 17.57          |
• A consensus between all amalgamated groups regarding the high relevance from the following characteristics of agility: being collaborative, adaptability and feedback incorporation;

• A consensus in all amalgamated groups regarding the high relevance of the agile practice continuous integration;

• The influence of some group skills on evaluating the relevance of some characteristic such as continuous testing, reflection and introspection, people oriented and practices such as test driven development, refactoring and collective code ownership;

• The low relevance of the agile practices on site customer and planning game and the characteristic of emergence in all amalgamated groups;

• Evidence on aggregating the results that participants from previous trials (original study) have a significantly different opinion regarding the relevance of the practices planning game, whole team and sustainable pace and the characteristic transparency;

• The reintroduction of the practice pair programming in the set of agile practices that had been previously discarded in the two first trials.

Although the benefits for samples’ representativeness observed applying LinkedIn as source of recruitment, it is important to highlight some relevant limitations on its use. It was necessary the use of “Premium” account for accessing so many subjects. Even using such account, LinkedIn restricts the number of retrieved results for members’ searching, retrieving in first place members more directed connected with the user’ account (de Mello et al. 2014a). However this behavior is not observed in groups of interest’ searching. To preserve the randomness on sampling, operational risks were introduced applying the additional filters mentioned in the subsection 5.3. Regarding the recruitment, it is important to emphasize that the user account need to be a member for each desired group of interest to send individual messages for its members. Also, one can see that the use of browser macros for sending such individual messages is error prone.

6 Conclusion

This paper exemplified through a survey on characteristics of agility and agile practices in software processes the use of a conceptual framework for supporting sampling in large scale SE surveys and how it can contribute for retrieving representative samples. As a result, a set of pertinent groups of interest from the professional social network LinkedIn could be stratified, suggesting similarities between them. From these strata, it was possible to randomly recruit a representative sample of 7745 individuals distributed in all continents. Then, based on the characterization reported by the respondents, the strata were reorganized into the following five amalgamated groups: agilists, testing professionals, programmers, configuration managers and system analysts. This process allowed us to identify relevant findings regarding the survey’s object of study.

Our experience with LinkedIn shows that it represents an interesting source of sampling for composing relevant sampling frames based on groups of interest, mainly whether the facilities associated with a “Premium” account are available for the
researchers. Besides, the use of automated tools to support the sampling and recruitment of LinkedIn members allowed us to mitigate operational errors, such as recruit a same subject twice or even bypassing someone else. Although SE doesn’t have yet specialized and widely adequate sources of sampling for subjects recruitment, we assert that it is possible to reduce the population bias and enlarge samples in SE surveys whether a systematic recruitment is planned over a professional social network and its results are analyzed through probabilistic sampling methods. As next step, we intend to extend the conceptual framework description, introducing guidelines for supporting its application and using it to support the conduction of more studies in Software Engineering.

**Abbreviations**

CI: Confidence interval; CL: Confidence level; CSV: Comma separated values; SE: Software engineering; SLR: Systematic literature review; SW: Software.

**Competing interests**

The authors declare that they have no competing interests.

**Authors’ contributions**

RMM carried out the planning, execution and data analysis of this research as part of his ongoing Doctoral Thesis and worked in the manuscript. PCS developed the automated support for the study and helped in the manuscript. GHT supervised the research and also worked in the data analysis and manuscript. All authors read and approved the final manuscript.

**Authors’ information**

Rafael M. de Mello is a D.Sc. student in the Experimental Software Engineering (ESE) group at COPPE/UFRJ. Pedro Correa da Silva is an undergraduate student at Polytechnic School/UFRJ and trainee researcher (UFRJ/PIBIC grant) at ESE group. Guilherme H. Travassos is a Professor of Software Engineering at COPPE/UFRJ and a CNPq 1D researcher. He leads the ESE Group since 2001.

**Acknowledgements**

The authors thank all subjects that took part in the studies and have collaborated with this research. CNPq has supported ESE group’s researches.

**Received: 10 December 2014 Accepted: 28 May 2015**

**Published online: 10 June 2015**

**References**

Abrantes JF, Travassos GH (2013) Towards Pertinent Characteristics of Agility and Agile Practices for Software Processes. CLEI Electronic Journal 16:1, Article No. 6, Montevideo, Uruguay

Basten D, Mellis W (2011) A Current Assessment of Software Development Effort Estimation. Proceedings of 4th International Symposium on Empirical Software Engineering and Measurement, 2011. IEEE, New York City, USA, pp 235–244

Bennett S, Woods T, Liyanage WM, Smith DL (1991) A simplified general method for cluster-sample surveys of health in developing countries. World Health Stat Q 44(3):98–106, Geneve, Switzerland

Bettenburg N, Just S, Schröter A, Weiss C, Premraj R, Zimmermann T (2008) What Makes a Good Bug Report? Proceedings of the 16th ACM SIGSOFT International Symposium on Foundations of Software Engineering, 2008. ACM, New York, pp 308–318

Chen L, Ali Babar M, Nuseibeh B (2013) Characterizing architecturally significant requirements. IEEE Softw 30(2):38–45. doi:10.1109/3-540-45143-3_7, New York, USA

Ciolkowski M, Laitenberger O, Vegas S, Briff S (2003) Practical experiences in the design and conduct of surveys in empirical software engineering. In: Wang A (ed) Conradi R. Empirical Methods and Studies in Software Engineering- Experiences from ESERNET, Springer Berlin Heidelberg, pp 104–128

Conradi R, Li J, Skyngstad OPN, Kampenes VB, Bunse C, Morisio M, Torchiano M (2005) Reflections on conducting an international survey of Software Engineering. Proceedings of 2005 International Symposium on Empirical Software Engineering. IEEE, New York, USA

de Mello RM, Travassos GH (2013) Would Sociable Software Engineers Observe Better? Proceedings of the 7th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2013. IEEE, New York, USA, pp 279–282

de Mello RM, da Silva PC, Travassos GH (2014a) Investigating Probabilistic Sampling Approaches for Large-Scale Surveys in Software Engineering. Proceedings of 11th Workshop on Experimental Software Engineering 2014, Pucón, Chile, 23–25 April 2014

de Mello RM, da Silva PC, Travassos GH (2014b) Sampling improvement in software engineering surveys. Proceedings of the 8th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2014. ACM, New York, USA, pp 13–17
