Investigating Probabilistic Sampling Approaches for Large-Scale Surveys in Software Engineering

Rafael Maiani de Mello, Pedro Correa da Silva, Guilherme Horta Travassos

Universidade Federal do Rio de Janeiro
PESC/COPPE
rmaiani@cos.ufrj.br
pedroorez@poli.ufrj.br
ght@cos.ufrj.br

Abstract. Surveys and quasi-experiments in software engineering are typically conducted over small populations defined by convenience, which hampers the generalization of observed evidence. In this sense, we argue that a strategy to support the systematic establishment of large population and correspondingly samples can contribute to reduce these threats. However, besides just proposing additional enlargement sampling alternatives, it is necessary to evaluate their use in actual Software Engineering studies. Therefore, this paper describes an experience regarding the recruitment of subjects concerned with the replication of a large scale survey on software processes agility characteristics through a professional social network (a.k.a. LinkedIn). By doing so, it was possible to deal with a sampling frame with more than 110,000 members distributed over 19 interest groups. Next, using similarity levels observed among the groups, eight strata were organized from which a sample with more than 7,700 subjects was able to be used to randomly recruit the participants. So far this replication obtained more participants and higher confidence level when compared with the two previous survey's trials with subjects recruited by convenience. Therefore, we believe the results from this experience, including the instruments and sampling plan can be useful for those researchers interested on executing large scale surveys in Software Engineering.

Keywords: random sampling, experimental software engineering, clustering, stratified sampling, survey, multivariate analysis, graph, strongly connected components.

1 Introduction

A major issue regarding the external validity of Software Engineering (SE) studies is concerned with the population usually defined by convenience and small samples [1,2,3]. As a consequence, the opportunities for replicating *quasi*-experiments and undertaking meta-analysis are limited [4], which hamper the evolving of SE practice [5]. A similar scenario can be observed in SE surveys, frequently used to gather the opinion of specialists and practitioners regarding SE technologies and practices [6,7,8]. Besides,

surveys are commonly used in quasi-experiments to support the collection of qualitative data, such as subjects' characterization and follow-up questionnaires [9].

Although the technical literature presents alternatives such as simulation [10,11] and crowdsourcing technologies [12,13] to enlarge sample's size, we argue that we also need to investigate realistic sets of SE subjects in order to bring more realistic results. In this sense, we recently conducted a pilot study [7] by applying a systematic recruitment plan to identify a large set of LinkedIn[1] members, composing the sampling frame for a survey on requirements effort estimation. The results obtained from this first experience has shown that the set of contributions extracted from this sampling frame is more heterogeneous and significantly more specialized in requirements engineering than the sample defined by convenience. Also, after this pilot execution, a set of lessons was compiled, directing our efforts to improve the development of new recruitment plans.

Currently, our experience has been based in the replication of a large scale survey regarding a set of agility characteristics and practices identified in a *quasi*-systematic literature review [8] and validated through a survey distributed by convenience. The referred survey was already executed twice always based on external researchers' opinion, identified in the list of papers included in the review. Considering both trials, the opinion of only 25 authors from the 158 recruited were collected and just 7 from them declared High or Very High experience applying agile approaches on software projects. Thus, in order to identify a more realistic and large-scale set of SE practitioners and researchers concerned with the survey's topic, as in our successful first experience, we decided to use LinkedIn as source of recruitment (SoR). Considering LinkedIn spreads across different areas of interest, the following research question started to direct our investigation: *"Who is the population from the SoR interested in Agility characteristics and practices concerned with Software Engineering?"*

Therefore, this paper describes the whole process of characterizing the population and the sampling frame in order to randomly invite LinkedIn members to answer a large scale survey on characteristics of agility in software processes. As far as we are aware, this is the first reported experience concerned with the application of sampling approaches studies based on social network populations in software engineering.

Section 2 describes the plan of recruitment, which execution is presented on section 3, where the sampling frame is presented. Section 4 analyses the feasibility of clustering or stratifying the referred sampling frame. Section 5 describes the automated support for the referred plan. Section 6 highlights a set of lessons learned during this investigation.

2 Plan of Recruitment

The following subsections describe the plan of recruitment designed for supporting the characterization of the survey population and for performing subjects' recruitment.

[1] http://www.linkedin.com

2.1 Search Criteria

The SoR offers facilities for filtering members and groups, having fewer restrictions for filtering groups. Thus, we defined *LinkedIn* groups as the search unities to be investigated, based on the following algorithm:

For each keyword, do:
Submit a keyword (between quotes) followed by the term "software" in the option "Group Search";
Identify all groups returned, recovering the following data: name, description, group rules and number of members.

2.2 Technical Terms (keywords)

Considering the set of agile practices and characteristics identified in the technical literature [8], it was defined the following set of keywords:
"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", "planning game"

2.3 Exclusion Criteria

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

- explicitly prohibits the execution of studies;
- explicitly restricts the individual messaging between its members;
- is explicitly directed to a city, region or country;
- is focused on promote 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;
- has a non-English language as default.

2.4 Recruitment Process

For each selected and included group from the SoR, it will be analyzed the coverage of this group over the entire population, in order to establish a sampling frame. For each group included on the sampling frame, the researcher with a "Premium" account in *LinkedIn* will subscribe on it, since only using this type of account it is possible to

identify members' overlapping between groups, supporting the definition of the net population size. Also, with this account, it is possible to recover members from a set of groups, not only from a single group, which is essential to support re-grouping analyzes. After the acceptance on the referred groups, it will be analyzed the feasibility of statistically re-grouping them to perform clustering and/or stratification sampling. If it is not possible to re-grouping, it will be applied Simple Random Sampling (SRS). After that, it will be calculated the size of the sample(s) according to the sampling approach adopted, randomly selecting subjects for the recruitment(s). The recruitment message will be individually sent using exclusively the message service provided by the SoR.

3 Execution

Executing the plan described on Section 2, we identified (September, 2013) 289 distinct groups, being excluded 227. Table 1 shows the incidence of each exclusion criteria. It is important to highlight that the 289 groups nor prohibits the execution of studies and nor restricts the sending of individual messages among their members, provided by LinkedIn by default. The sum of incidences on Table 1 extrapolates the total number of groups excluded since some groups were excluded due to two or more reasons.

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%**

We analyzed the size of each 62 included groups, distributing them over quartiles, descending ordered by the number of members, as presented on Table 2. It was observed that more than 92% of the gross population (including the repetition of members on each group), are presented only in the third quartile. Thus, we decided to consider only the third quartile (16 groups) on the research, establishing the first version of the sampling frame. However, it is important to highlight that to recruit member from groups on the SoR, it is necessary the acceptance of the researcher account on each group. As after a month since the subscriptions the researcher was not accepted on the groups G and O, we substituted them by the groups Q, R, S, T and V, where the same researcher was immediately accepted. By this way, it was maintained 90% from coverage. Thus, Table 3 presents the final version of the sampling frame. Although some included groups are not explicitly related with the agile context, they were not excluded due our intention in following the recruitment plan.

Table 2. Distribution of selected Groups in quartiles.

Third Quartile				Second Quartile				First Quartile			
Grp	**Size**	**Grp**	**Size**	**Grp**	**Size**	**Grp**	**Size**	**Grp**	**Size**	**Grp**	**Size**
A	47678	I	8225	Q	5464	Y	421	AF	201	AN	26
B	39770	J	7666	R	4043	Z	399	AG	124	AO	24
C	37116	K	7324	S	3780	AA	373	AH	116	AP	21
D	26783	L	6931	T	2359	AB	258	AI	104	AQ	19
E	21009	M	5961	U	1747	AC	258	AJ	51	AR	19
F	20967	N	5710	V	913	AD	237	AK	45	AS	17
G	13958	O	5690	W	747	AE	235	AL	45	AT	17
H	9940	P	5653	X	731			AM	28		

Table 3. Groups included on the sampling frame.

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

In November 2013, the sampling frame totalized 264,540 gross members, distributed over 149 countries. Using filtering resources from the SoR, we identified 202,643 distinct members, i.e., an overlapping rate of 30.54%. Also, we identified a great geographical concentration (76.25%) at only ten most populous countries: USA (35.70%), India (15.15%), UK (9.41), Canada (4.07%), Australia (3.07%), Brazil (2.35%), Netherlands (2.33%), Sweden (1.88%), France (1.71%) and China (1.25%). Table 4 presents the distribution of distinct members over great regions of the world, considering the 60 more representative countries (98.23% from total).

Table 4. Distribution of distinct members subscribed on the sampling frame groups, considering the 60 most represented countries.

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

4 Grouping Analysis

Since it was established the sampling frame, the next step includes identifying the sampling design to be applied. In order to avoid selection biases, it was considered only probabilistic sampling designs, in which each unity from the sampling frame has the same chance of being selected. The most common probabilistic sampling approach is the simple random sampling (SRS), where all the unities from the sampling frame have the same probability to be recruited [14]. However, SRS don't take into account the similarities between sets of units. Thus, considering the sampling frame is initially represented by 19 *LinkedIn* groups, next subsections discuss the feasibility of applying another two probabilistic sampling approaches: *clustering* and *stratification*.

4.1 Clustering

Clustering is a sampling approach commonly used as alternative to reduce the efforts of recruitment, without reducing the whole population representation. The concept of clustering sustains that sets of unities from a population having a great level of similarity composes a single cluster, making feasible to apply SRS over it [14], i.e., randomly discarding one or more sets from the sampling process. This approach is frequently applied on medicine [15,16,17], especially when it is necessary access risk areas to perform the observations.

In the case of the presented study, due to the limitations from the SoR to recover the detailed and updated profile from its members, we considered to calculate the similarity levels as the overlapping rate between the 19 groups, i.e., if a group "X" has an overlapping distribution with other 18 groups similar with the distribution of a "group Y", we could amalgamate them into a single cluster. Thus, we calculated the overlapping rate (S) from each group "i" over each group "j", according to the following formula (1):

$$Sim_{ij} = ((PG_i + PG_j) - PG_i)/ PN_{ij} \quad (1),$$

where PG is the total number of members in a single group (gross size) and PN is total number of distinct members between two groups (net size).Figure 1 shows an extract from the resultant matrix of overlapping. Based on this matrix, it was conducted the multivariate analysis, supported by the resource "cluster observations" available on MiniTab[2]. In our analysis, we applied the "Average" linkage and Pearson's correlation to calculate distances between clusters. Thus, the dendogram showed on the Figure 2 represents the points where the clusters were performed until performing a single cluster.

Considering the great similarity between variables necessary to discard one or more sets, we established the minimum level of 75% to perform a cluster. However, as could be observed on Figure 2, we just found amalgamations performed under this level. In fact, the best level of similarity was found between the groups "H" and "I" (50.35%). Thus, based on these observations, we concluded that the clustering approach for the referred sampling frame is unfeasible, discarding this alternative for the current study.

	A	B	C	D	E	F	H	I
A	100.00%	4.22%	25.95%	27.60%	25.42%	3.91%	4.35%	3.77%
B	3.48%	100.00%	3.43%	4.00%	2.25%	24.19%	3.01%	2.33%
C	20.04%	3.21%	100.00%	18.65%	20.13%	2.97%	2.62%	2.20%
D	15.49%	2.73%	13.56%	100.00%	16.38%	2.71%	2.35%	2.39%
E	11.32%	1.21%	11.61%	12.99%	100.00%	1.10%	1.52%	1.23%
F	1.70%	12.72%	1.66%	2.09%	1.07%	100.00%	1.62%	1.91%
H	0.90%	0.75%	0.70%	0.87%	0.71%	0.77%	100.00%	45.66%
I	0.64%	0.48%	0.49%	0.73%	0.47%	0.76%	37.73%	100.00%

Fig.1. Extract from the matrix of overlapping.

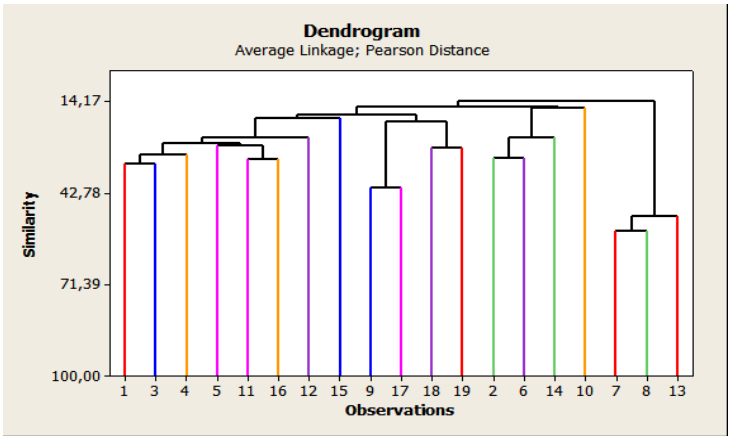

Fig.2. Dendogram referred to the clustering of 19 groups from the sampling frame.

4.2 Stratification

Stratification is a statistical sampling approach that takes into account the whole population heterogeneity, performing the *strata* in which independent analyses shall be performed. When the *strata* are well established, we have homogeneous sub-populations independently sampled, bringing more precise results than SRS [14].As we didn't find in the technical literature methods for stratification in similar contexts (social networks and software engineering), we decided to verify its feasibility through two distinct approaches: the multivariate analysis and the analysis by digraphs, aiming at identifying strongly connected components.

4.2.1 Multivariate Analysis

Although stratification is commonly used to organize subjects under non-overlapped strata, such as geographical regions [14], in the context of social networks groups, the overlapping of subjects is expected and could suggests affinity between them, supporting the re-organization of the sampling frame into a set of "groups of interest". Based on the clustering trial mentioned on section 4.1, we observed an initial amalgamation

behavior between groups 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, we established a reduced minimum similarity level (25%) for amalgamating groups into one stratum. Based on this threshold, we identified seven strata. Table 5 presents each stratum and its composition.

Table 5. The seven strata derived from multivariate analysis.

Stratum	Name	Groups
EM1	Agility	A, C, D, E, L, R, M
EM2	Project Management	Q
EM3	Agile Practices 1	J, S
EM4	Agile Practices 2	T, U
EM5	Software Testing	B, F, P
EM6	Configuration Management	H, I, N
EM7	Software Architecture	K

4.2.2 Digraph Analysis

As an alternative for the analysis presented on section 4.2.1, we performed a digraph analysis, in order to identify strong connections between the 19 groups (two-by-two). For this, it was considered each group as a node and the weight of each directed connection as the already calculated overlapping rate. Thus, similarity in this case has been observed by considering the connectivity between couple of groups (two-by-two). However, considering that all 19 groups are connected with each other in both directions, we decided to filter only the greatest rates. Thus, all the 342 overlapping rates (minimum 0.02% and maximum 48.67%) were distributed by quartiles. Thus, considering low rate of all quartiles (0.56%, 1.98% and 6.14%), we decided to represent on the digraph only the rates presented on the third quartile (6.14%), resulting on a digraph with 83 connections.

Then, it was performed the Strongly Connected Component (SCC) algorithm, available in the GraphViz[2] tool. As consequence, the original digraph was disjoint into five components, one digraph with 11 nodes (A, C, D, E, L, R, M, J, S, T, U), a second one including three nodes (H, I and N, digraph "b" on Figure 3), and a third digraph with two nodes (B and F, digraph "a" on Figure 3). In addition, three other nodes maintained disconnected with other nodes: K, P and Q. Thus, based on this approach, we identified the suggestion of establishing the six strata presented on Table 6.

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

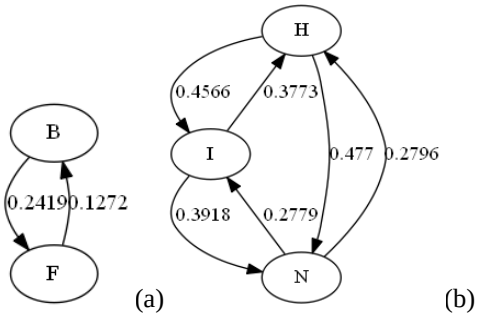

Fig.3. Two digraphs derived from the original digraph with 19 nodes.

Table 6.The six strata derived from digraph analysis.

Stratum	Name	Groups
ED1	Agility and Agile Practices	A, C, D, E, L, R, M, J, S, T, U
ED2	Project Management	Q
ED3	Software Testing 1	B, F
ED4	Software Testing 2	P
ED5	Configuration Management	H, I, N
ED6	Software Architecture	K

4.2.3 Comparison between analyses

Making a comparison between the analyses presented on the two last sections, one can see that although the digraph analysis suggests fewer groups, it presents various similarities with the multivariate results. As main differences, it could be observed that digraph analysis unifies into a single stratum (ED1) three strata referred to agility (EM1) and agile practices (EM3 and EM4). At the same time, the digraph analysis divided the stratum from multivariate analysis related with software testing (EM3) into two strata (ED3 and ED4). Thus, aiming to avoid overestimated amalgamations, it was chosen preserving all the identical strata in both analysis and selecting only the different strata having fewer groups. The result is the set of strata presented on the Table 7.

Table 7. The final stratification proposed.

Str.	Name	Groups	Str.	Name	Groups
E1	Agility	A, C, D, E, L, R, M	E5	Software Testing 1	B, F
E2	Proj. Management	Q	E6	Software Testing 2	P
E3	Agile Practices 1	J, S	E7	Conf. Management	H, I, N
E4	Agile Practices 2	T, U	E8	Software Architecture	K

5 Sampling and Recruitment

Considering the performing of stratification sampling, we need to compute independent sample sizes for each stratum, attending to the same confidence interval and the same confidence level. Thus, considering the number of members from each stratum and the technical limitations from the SoR to randomly recover group members, we defined a sample size based on the level of confidence of 99% and on a confidence interval of 3.5%, which corresponded to the sample sizes showed on Table 10.

However, the *LinkedIn* users, still having "Premium" accounts, can't retrieve more than 700 members from each group. Thus, the default search into groups was sufficient only to retrieve samples with distinct members for the strata E1, E4 and E7. Thus, aiming at to enlarge the number of members retrieved for the other strata (E2, E3, E5, E6 and E8), we added a search filter (members with at least 10 years of professional experience/ members with less than 10 years of professional experience), potentially retrieving 1,400 members. At the end, Table 8 shows that it was retrieved at least 10% more members than the sample size needed for each stratum.

Table 8. Sample size for each stratum and the respective number of member retrieved.

Str.	#Distinct Members	Sample Size	Retrieved Members	Str.	#Distinct Members	Sample Size	Retrieved Members
E1	114,827	1,031	2,563	E5	56,400	1,021	1,358
E2	5,488	874	1,401	E6	5,791	882	1,391
E3	11,633	955	1,444	E7	17,234	981	1,016
E4	3,864	820	883	E8	7,335	911	1,438

5.1 Retrieved Members Analysis

It is important to highlight that the SoR has a biased behavior on searching, preferentially retrieving the members attending to the search criteria, but also 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 discarding these members, since they are part from the population, we included them on the strata and, at the same time, evaluated the impact derived from this bias. For this evaluation, we considered the 10 more representative countries over the whole population. Tables 9 and 10 show the distribution of the sampling frame population between these countries and the retrieved members' distribution between the same countries, respectively. Based on both distributions, Table 11 shows the correlations calculated for each country, considering a confidence coefficient equals to 95%. These correlations were calculated in order to observe how much the country distribution from the retrieved sample is similar to the sampling frame, suggesting a good sample.

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 "parentage" over the "bad" correlations obtained for E1, E3 and E5, it was observed that this 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 "parentage". Thus, we experimented removing this set of "parents" from these strata. However, when recalculated their respective correlations, we identified that only E1 was significantly influenced by this removing with the increasing of its correlation from to 0.9745. Thus, we decided to maintain the referred change only to E1.

Table 9. Sampling frame members by ten most represented countries.

Str.	USA	IND	UK	CAN	AUS	BRA	NET	SWE	FRA	CHI
1	37557	9337	11346	4355	3878	3228	3065	2509	2305	971
2	1547	535	471	274	238	199	131	115	117	90
3	2236	1811	2253	351	488	486	374	244	322	74
4	939	328	406	177	89	155	108	77	77	38
5	18692	14942	3743	2115	1192	606	814	644	456	941
6	3132	1069	235	132	43	25	20	57	16	245
7	7723	2902	1501	683	407	155	316	353	250	202
8	2371	1129	353	323	120	160	280	67	128	79

Table 10. Retrieved members from the ten most represented countries.

Str.	USA	IND	UK	CAN	AUS	BRA	NET	SWE	FRA	CHI
1	1016	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 11. Correlation between the distributions showed in Table 11 and 12.

Str.	Correlation	95% (inferior)	95% (superior)	Str.	Correlation	95% (inferior)	95% (superior)
E1	0.8022	0.3487	0.9513	E5	0.7611	0.2524	0.9402
E2	0.9957	0.9814	0.999	E6	0.9786	0.9092	0.9951
E3	0.7857	0.309	0.9469	E7	0.943	0.7714	0.9868
E4	0.9595	0.8333	0.9906	E8	0.9843	0.9329	0.9964

5.2 Recruitment

Based on the final list of members retrieved from each stratum, the members were randomly ordered using the tool from Random.org[3] and it was selected the "n" first distinct members from each stratum, aiming at obtaining the expected sample size, as described

[3] http://www.random.org

in Table 10. Due to the possibility of finding the same member in distinct strata, the selection occurred from the stratum with less surplus (E7) to the higher excess of retrieved members (E1). At the end, all the 7,745 distinct expected members were invited through LinkedIn messages.

5.3 Automated support

In order to make feasible selecting and sending a large amount of invitations using the SoR, we automated two activities: the data extraction from the users profile and the message sending through the social network. To extract the data required to select the members we used a free extension available for some browsers named iMacros[4]. This extension makes possible to select HTML tags from a webpage, extracting its content. iMacros also allows creating macros using its own scripting language. Combining these two features, we developed two macros: one to browse the user profiles from a group extracting their data and another to allow the sending of invitations through the social network mail service.

The first macro loops across each listed member on the search results pages, selecting its name, connection level, country, job and its profile URL. At the end, a CSV file stored all the retrieved information from the members. Then, we implemented a Ruby script to refine this file, removing any extra information and extracting the User Id and Group Id from the profile URL. This new file fed a PostgreSQL[5] database, supporting several observations made on this research.

After selecting the users to be recruited, as mentioned on the last subsection, it was listed in another CSV file the user id and the group id for which the invitation should be sent. The second macro takes action using this CSV file to generate personal URLs and a default message, composing a sending invitation for each selected members personal messages for each member. The use of automated support can introduce a set of threats to validity mainly due to the use of macros extracting text from dynamic webpages. However, we believe it does not invalidate our proposed strategy.

6 Lessons Learned

Considering the plan, the source of recruitment and the sampling approaches analyzed in the previous sections, we identified a set of important lessons for future studies. First, we observed the positive impact of using the results from a previous systematic review to support the group searching, bringing an unbiased and diverse population. However, it is not clear how *LinkedIn* implements its group searching.

Also, for the referred research theme, we observed that a set of few LinkedIn groups concentrate most of members, providing a significant reduction of complexity for establishing the sampling frame. In addition, it was observed that the biggest groups are more driven to the discussion of technologies and practices than to supporting headhunting and job seeking activities, involving practitioners from all seniority levels

[4] http://wiki.imacros.net
[5] http://www.postgresql.org

and researchers. Thus, it reinforces the idea of *LinkedIn* as a good source to improve the quality and size of samples, as previously observed [12].

Considering the analyzed sampling approaches, we highlight the observed feasibility of stratifying social network groups, considering how its members are distributed over the other groups on the sampling frame (overlapping). Also, it was observed a positive relation between the affinity of group themes and the overlapping rates observed between their members, suggesting a trend from LinkedIn members on subscribing in specific groups of interest. However, it was not found sufficient similarity to perform clusters, suggesting also the heterogeneity between the investigated groups. Finally, despite all the already mentioned technical limitations, it was observed the feasibility of randomly recruit significantly and well distributed samples for the referred survey.

7 Conclusion

This paper presented a systematic plan developed to support the establishment of population and its sampling to answer a large scale survey on characteristics of agility in software processes. The results derived from the execution of this plan brought us a set of groups that could be stratified, suggesting similarities between them and, at the same time, the heterogeneity from the sampling frame. From these strata, it was possible to recruit 7,745 potentially software engineering researchers and practitioners, which could be considered a representative set from the whole population concerned with the research theme. Also, until the submission of this paper, we had received 280 answers before 11 days of execution, i.e. a confidence level (Hamburg) of 94.1% from the whole sample. As a comparison with the previous trials of the same survey, on the first trial it was recruited 113 subjects and it was obtained only four answers; on the second trial, 117 subjects were recruited and 21 answered, totalizing confidence levels of 50.9% and 80.2%, respectively. Therefore, in order to find similar results on further SE studies, we recommend to design specific recruitment plans describing all the four main steps presented in this paper (section 2): define a searching criteria considering peculiarities from the SoR; define a set of keywords based on the technical literature; identify a set of exclusion criteria; describes how the sampling and the recruitment will be performed, only considering probabilistic sampling approaches. As source of recruitment, we observed *LinkedIn* as a good alternative since there is a "Premium" account available. Also, the use of tools to support sampling and recruitment, can avoid operational errors, such as recruit the same subject twice and calculating wrong the sampling frame size.

Although SE don't have yet specialized and widely representative sources for subjects recruitment, we assert that it is possible to reduce the population bias and enlarge sampling in SE surveys when the systematic recruitment is planed over a professional social network and its results analyzed through probabilistic sampling methods. As next step, we will analyze the results obtained after the recruitment. Also, considering that we already developed systematic plan over three distinct surveys (one of them yet not published), we intend to apply the same strategy to support the replication of *quasi*-experiments, being able to compare the results (when available) with recruitments made by convenience.

Acknowledgements

We thank CNPq for supporting this research.

References

1. D. I. Sjøberg et al. “A survey of controlled experiments in software engineering.”, IEEE Transactions on Software Engineering, vol. 31(9), pp. 733-753, 2005.
2. T. Dybå, V. B. Kampenes and D. I. K. Sjøberg. “A systematic review of statistical power in software engineering experiments,” Inf. and Soft. Tech., vol. 48, pp. 745–755, 2007.
3. L. M. Pickard, B. A. Kitchenham and P. W. Jones “Combining empirical results in software Engineering,” Information and Software Technology, vol. 40, pp. 811–821, 1998.
4. B. A. Kitchenham, et al. “Systematic literature reviews in software engineering – A tertiary study,” Information and Software Technology, vol. 52, pp. 792–805, 2010.
5. R. M. de Mello and G. H. Travassos. “An Ecological Perspective Towards the Evolution of Quantitative Studies in Software Engineering,” Proc. of 17th EASE, pp. 216-219, 2013.
6. G. Santos et al. “MPS. BR Program and MPS Model: Main Results, Benefits and Beneficiaries of Software Process Improvement in Brazil”. Proc. of Intl. Conf. on Quality of Information and Communications Technology (QUATIC), pp. 137-142. IEEE, 2012.
7. R. M. de Mello and G. H. Travassos “Would Sociable Software Engineers Observe Better?” Proc. of 7th ESEM, 2013.
8. J. F. Abrantes and G. H. Travassos. “Towards Pertinent Characteristics of Agility and Agile Practices for Software Processes”. CLEI Electronic Journal, vol. 16(1), 2013.
9. A. M. Fernández-Sáez et al.“Are forward designed or reverse-engineered UML diagrams more helpful for code maintenance?: a controlled experiment,” Proc. of the 17th EASE, ACM, 2013.
10. D. M. Raffo and M. I. Kelner. “Empirical analysis in software process simulation modeling,” The Journal of Systems and Software, vol. 53, pp. 31-41, 2000.
11. R. E. Garcia, M. C. de Oliveira and J. C. Maldonado. “Genetic algorithms to support software engineering experimentation,” Proc. of Intl. Symp. on Empirical Soft. Eng., 2005.
12. K. T. Stolee and S. Elbaum. “Exploring the Use of Crowdsourcing to Support Empirical Studies in Software Engineering,” Proc. of 4th ESEM, 2010.
13. K. T. Stolee and S. Elbaum. “On the Use of Input/Output Queries for Code Search” Proc. of 7th ESEM 2013.
14. S. K. Thompson. “Sampling,” 3rd edition, John Wiley & Sons, 2012.
15. S. Bennett et al. “A simplified general method for cluster-sample surveys of health in developing countries.” World Health Stat Q vol. 44(3), pp. 98-106, 1991.
16. L. Roberts et al. “Mortality before and after the 2003 invasion of Iraq: cluster sample survey.” The Lancet vol. 364 (9448), pp. 1857-1864, 2004.
17. S. M. Eldridge, D. Ashby and S. Kerry. “Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method.” International Journal of Epidemiology vol. 35(5), pp. 1292-1300, 2006.