Towards a Framework to Support Large Scale Sampling in Software Engineering Surveys

Rafael Maiani de Mello, Pedro Correa da Silva
COPPE/PESC
Federal University of Rio de Janeiro
P.O. Box 68511, Brazil
+55 21 2562 8654
rmaiani@cos.ufrj.br,
pedroorez@poli.ufrj.br

Per Runeson
Department of Computer Science
Lund University, Sweden
per.runeson@cs.lth.se

Guilherme Horta Travassos
COPPE/PESC
Federal University of Rio de Janeiro
P.O. Box 68511, Brazil
+55 21 2562 8654
ght@cos.ufrj.br

ABSTRACT
Context: The low quality and small size of samples in empirical studies in software engineering hamper the interpretation and generalization of their results. Therefore, enlarging sample sizes and improving their quality represent an important research challenge. **Goal**: We aim to define a conceptual framework, including requirements for establishing adequate sources for sampling subjects in software engineering surveys. **Method**: We use previous experience on applying systematic sampling strategies combined with contemporary web technologies in previously executed surveys, to organize the conceptual framework. We analyze its application to different sources of sampling. **Results**: The framework was observed to be feasible after its application to nine different large-scale sources of sampling. **Conclusions**: The analyzed crowdsourcing tools do not support essential requirements to be considered sources of sampling, while freelancing tools and professional social networks do.

Categories and Subject Descriptors
D.2.4 [Software Engineering]: Software/ Program Verification–statistical methods.

General Terms
Experimentation, Human Factors.

Keywords
experimental software engineering; sampling; population; quantitative studies; survey; sampling frame.

1. INTRODUCTION
To find the right population for empirical studies is a continuous challenge in Software Engineering (SE). The human subjects for surveys and experiments are usually selected through sampling by convenience [1,2,3]. The quality of the sample depends on how the population is established and to which extent its representativeness allows the extraction of meaningful samples. In this context, Pickard et al. [4] argue that without protocols to establish populations in SE and a set of standard measures recorded in experimental studies, meta-analysis becomes unfeasible. In addition, one can see that the lack of sources for retrieving adequate samples in SE contributes to the current scenario of reduced strength of the empirical evidence [5].

Thus, our research aims to contribute to improve population sampling, by depicting a *framework,* composed by a set of concepts and processes, aimed at supporting researchers on establishing adequate populations and samples for SE surveys. It is expected that the concepts organized into this framework can be applied in compliance with each study protocol, contributing to the improvement of the quality of the study. Besides, such concepts intend to stimulate the adoption of alternative sources of sampling, especially those based on Web technologies and social networks principles.

The sample quality can be assessed in terms of its *size*, *heterogeneity* and *confidence level* of its subjects, considering the context of each empirical study. So, as a first step towards a framework to support large scale sampling in SE surveys, we conducted a set of preliminary studies, addressing the following research question: *"Do systematic strategies for large scale recruitment of subjects in Software Engineering surveys contribute for improving the samples' quality?"* The preliminary studies were conducted on three surveys regarding different SE contexts, previously executed by the Experimental Software Engineering Group (COPPE/UFRJ), of which two are published [6, 7]. These surveys were replicated, changing their original sampling approach (by convenience) into random sampling on professional social networks, through systematic sampling strategies. In general, these strategies were designed aiming at supporting the main concepts regarding surveys population and sampling, available in statistics. The results [6] allow us to gather evidence regarding the usefulness of the applied sampling strategies with respect to improvement of the sample quality.

Then, based on the lessons learned, we organized a conceptual framework presented in the Section 2. Section 3 exemplifies the use of the presented framework concepts in order to evaluate a set of alternative sources of sampling available on the Web and based on contemporary technologies such as *freelancing* and *social networking* that can represent interesting candidates for large-scale sampling in SE research.

2. The Conceptual Framework
The framework concepts are presented in Figure 1, expressed in conceptual *class diagrams* (in white), associated with typical SE empirical studies components (in grey). It is based on four main concepts presented in the following subsections: *search unit, source of sampling, search plan and sampling strategy*.

ESEM'14, September 18–19, 2014, Torino, Italy.

http://dx.doi.org/10.1145/2652524.2652567

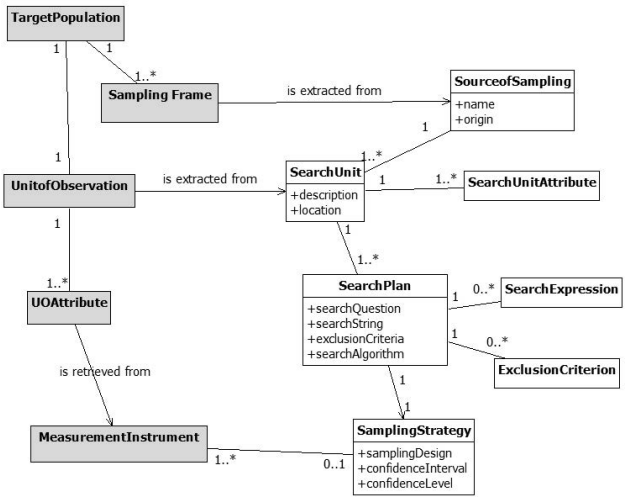

Figure 1. The conceptual framework concepts.

2.1 Search Unit

A *search unit* characterizes how one or more *units of observation* [8] can be retrieved from a specific source of subject recruitment. Ideally, search units must have a one-to-one correspondence with the units of observation. For example, surveying company use of a certain practice, one should search for companies, not professionals working in companies. However, the SE researcher must be able to deal with sources' limitations from where the samples can be extracted, such as exemplified in Figure 2.

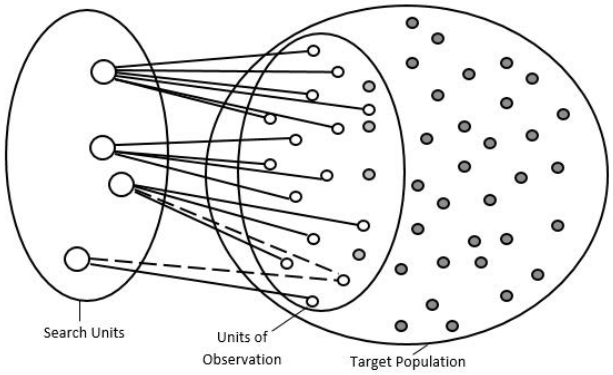

Figure 2. An example of n-to-m correspondence between search units and units of observation.

In this scenario, the retrieved search units can be composed by one or more units of observation. In addition, a single unit of observation can be referenced by two distinct search units (marked in dashed lines). These limitations can imply in greater effort on mitigating operational risks, but we argue that they do not invalidate the sampling process at all. In fact, the researcher must consider avoiding to recruit the same unit more than once, and working on clearly distinguishing each search unit, updating the available data when needed.

The scenario represented in Figure 1 can be observed, for instance, when the search units are composed by LinkedIn groups, but the units of observation are composed by their members, that could be found subscribed in one or more groups at the same time [6, 7]. Alternatively, in the study presented by Frakes and Fox [9], although the units of observation were established as software organizations, i.e., groups of practitioners, the search units were defined as software practitioners, without taking into account their respective organizations.

2.2 Source of Sampling

A Source of Sampling (SoS) consists of a database (automated or not) from which adequate *subpopulations* of the *target population* can be *systematically retrieved* and *randomly sampled*. To be considered valid, a SoS *candidate* shall satisfy the following *essential requirements* (ER):

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

Figure 3 suggests a set of *SoS candidates,* most of them extracted from the SE specialized literature, considering their applicability according to the specificity of the target population. However, it is important to highlight that a SoS candidate can be composed by more than one type of search unit. In this case, the researcher must choose the suitable search unit for each research context.

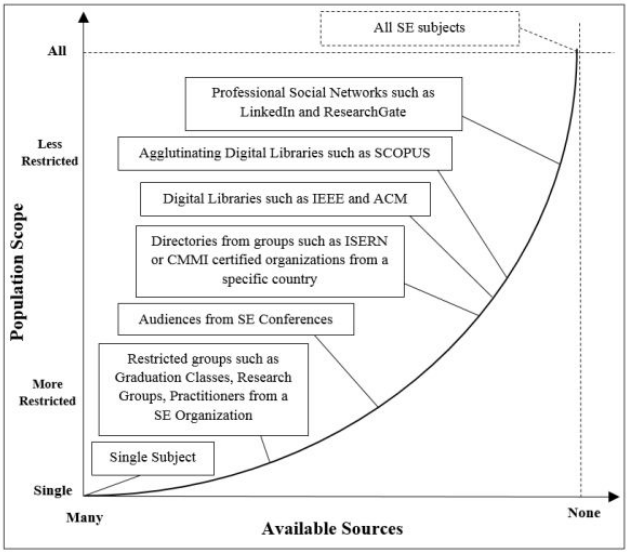

Figure 3. Examples of SoS candidates

This clearly indicates the key role of the target population on evaluating the validity of a SoS candidate. We argue that the balance between the target population and selected SoS is one of the main tasks when planning surveys and experiments. Without this balance, researchers cannot interpret the generality of the results. In addition, it is expected that an SoS also satisfies the following desirable requirements (DR), concerned with the samples' accuracy (ADR), clearness (CDR) and completeness (CoDR):

- *ADR1. It is possible to retrieve each search unit from the SoS in a logical and systematic way*.
- *ADR2. There are no units of observation outside the target population concerned with the SoS.*
- *ADR3. There is a one-to-one correspondence between each search unit and each unit of observation of the target population*.
- *CDR1. All search units appear once in the SoS.*
- *CDR2. All units of observation appear once in the SoS.*
- *CoDR1. All information needed from each search unit is up-to-date*.
- *CoDR2. All information needed from each unit of observation is accessible and up-to-date*.
- *CoDR3. All units of observation from the target population can be found in the SoS.*
- *CoDR4. Each search unit provides relevant information for performing alternative probabilistic sampling designs, such as stratified and cluster sampling. The SoS directly supports the researcher on classifying and/or clustering its population.*

Some requirements presented in this section include conditions for establishing ideal sampling frames as suggested by Särndal et al. [10], adding criteria addressed for dealing with the limitations of the SoS typically available for SE research. In statistics, a *sampling frame* is the source from which a sample is drawn, identifying all the units of a population that can participate in the sampling process and may include units such as individuals, households and institutions [10]. Thus, if a SoS can be considered valid for a specific research context, we can conclude that sampling frames can be established from this SoS for the same research context.

2.3 Search Plan

A *search plan* describes how *search units* will be systematically retrieved from a *SoS* and evaluated in order to be included or not into a *sampling frame*. The following subsections present the main components of a search plan. Examples of applying each component, are presented by de Mello *et al* [6,7].

2.3.1 Search String

A *search string* is composed by a set of *search expressions* connected through logical operators that can be applied to a SoS in order to retrieve search units. As in the case of systematic literature reviews [11], we argue that search expressions can be applied to make an unbiased filtering of units feasible. However, search strings must be avoided when the requirement ADR2 is satisfied. This could happen, e.g., when the SoS is composed of the list of employees from a SE organization, and the set of employees from this organization composes the target population.

2.3.2 Search Algorithm

The search algorithm describes each step, automated or not, that must be followed in order to filter the *search units* in a SoS, including how to apply the planned *search string*. A search algorithm can vary significantly in its complexity, depending on available SoS resources. In addition, any previously known restriction for accessing the search units must be described.

2.3.3 Exclusion Criteria

Another concept borrowed from systematic literature reviews, the *exclusion criteria,* describes a set of restrictions that must be applied in order to exclude undesirable search units retrieved from the search plan execution. Exclusion criteria can be especially helpful when the SoS is significantly generic, such as in the case of the professional social network applied in our preliminary studies [6,7]. However, as in the case of *search strings*, if the requirement ADR2 is satisfied, the establishment of exclusion criteria must be avoided.

2.4 Sampling Strategy

A *sampling strategy* describes the steps that must be followed in order to sample and access the units of observation that will be investigated in the study trial. A sampling strategy can be composed by the following attributes:

- the description of the probabilistic *sampling design* that will be applied (simple random sampling, clustering, stratified sampling, systematic sampling, etc…), when the full sampling is undesirable or unfeasible;
- the minimal confidence interval of the sample and the confidence level to support the calculating of the sample sizes;
- additional *measurement instruments* [8], such as questionnaires, in order to retrieve the set of data unavailable or not updated, in the case of the requirements CoDR1 and CoDR2 are not satisfied by the SoS.

3. ALTERNATIVE SoS

During our investigations, we observed some sources available on the Web that can be considered SoS candidates to support large-scale SE surveys. We present evaluations of these SoS considering the generic target population: "SE practitioners and/or researchers". In order to perform these evaluations, each SoS candidate was submitted to the essential and desirable requirements for a SoS presented in subsection 2. In total, we evaluated nine sources, grouped in the next subsections by three main types of Web-based technologies: *professional social networks, crowdsourcing tools* and *freelancing tools.*

3.1 Professional Social Networks

We consider a *professional social network* as a social network, having as its main goal to support the establishment of professional connections between individuals through resources such as technical forums, research dissemination, job advertisements and skill endorsement. In this context, we highlight three main SoS candidates: *LinkedIn* (L), *Academia.edu* (AC) and *ResearchGate* (RG). Table 1 synthetizes the evaluation of each source through the requirements for a SoS presented in the Section 2. In the case of *LinkedIn*, we performed two distinct evaluations, one for each identified search unit: *members* (LM) and *groups* (LG).

Table 1. Evaluation of Professional Social Networks

Src.	ER				ADR			CDR		CoDR			
	1	2	3	4	1	2	3	1	2	1	2	3	4
LG	Y	Y	Y	Y	Y	N	N	Y	N	N	N	N	N
LM	*Y*	*Y*	*Y*	*N*	-	-	-	-	-	-	-	-	-
RG	Y	Y	Y	Y	Y	N	Y	Y	Y	N	N	N	Y
AC	Y	Y	Y	Y	Y	N	Y	Y	Y	N	N	N	Y

As can be seen in Table 1, the *LinkedIn* sampling through members (LM) proved as an inadequate alternative of SoS since the network does not allow researchers to access all retrieved members' profiles, unsatisfying ER4. In the other hand, it is possible to access all *LinkedIn* retrieved groups (LG), as described in our preliminary studies [6, 7]. In the case of RG, although it was observed as an adequate SoS, we experienced considerable

limitations in practice, imposed by the network on performing the recruitment. Although it was observed that AC presents a similar behavior on accessing researchers' profiles to RG, we cannot compare their recruitment support since we did not use AC to support a survey.

3.2 Crowdsourcing Tools

We analyzed three distinct crowdsourcing tools designed for hiring professionals interested on performing paid tasks on many knowledge areas: *Mechanical Turk* (MT), *MicroWorkers* (MW) and *ClickWorkers* (CW). Table 2 resumes these evaluations considering *member* as the search unit. One can see that no analyzed tools satisfied all the essential requirements to be a SoS.

Table 2. Evaluation of Crowdsourcing Tools

Source	ER				ADR			CDR		CoDR			
	1	2	3	4	1	2	3	1	2	1	2	3	4
MT	*N*	*Y*	*Y*	*N*	-	-	-	-	-	-	-	-	-
MW	*Y*	*N*	*Y*	*N*	-	-	-	-	-	-	-	-	-
CW	*Y*	*Y*	*Y*	*N*	-	-	-	-	-	-	-	-	-

In fact, although these tools can be useful for improving the samples' size in SE studies, as already demonstrated in SE with MT [12, 13], we observed that MT, MW and CW do not allow the analysis of available populations hidden in the "crowd". Thus, they do not satisfy ER4. In addition, we identified with MT that the tool restricts the subscription for a small set of countries and with MW that the tool may bias the recruitment process, prioritizing the invitation of "better qualified" profiles.

3.3 Freelancing Tools

We have also identified a set of web tools that we named *freelancing tools*, where professionals can be hired to perform paid tasks. However, in contrast to the crowdsourcing tools, the user can filter, select and keep in touch with the professionals that he/she wants to hire. Thus, we find that this technology combines the advantages of the professional social network on filtering and accessing members' profiles with the possibility of creating tasks and the quality evaluation of each contribution from the crowdsourcing tools. Table 3 summarizes the evaluation of three freelancing tools: *e-Lance (EL), ODesk (OD) and Freelancer (FL)*, considering their *members* as search units.

Table 3. Evaluation of Freelancing Tools

Src.	ER				ADR			CDR		CoDR			
	1	2	3	4	1	2	1	2	3	1	2	3	4
EL	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	N	Y
OD	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	N	Y
FL	Y	Y	Y	Y	Y	N	Y	Y	Y	Y	Y	N	Y

Considering the interest of these tool members to maintain their profiles in order to be hired, and the concern of each tool with the hiring process, we infer that each profile tends to be continuously updated and unique. Also, it is important to highlight that these tools offer mechanisms to avoid the payment of invalid contributions. However, one can see that, in practice, the recruitments performed through freelancing tasks can be hard to be financially supported in large scale, since workers are paid by working hours.

4. CONCLUSION

The extent in which the threats to external validity can be mitigated in SE empirical studies is closely related with the extent in which sampling bias is avoided. Considering the lack of adequate sources for sampling in SE, especially in the case of large-scale studies, this paper introduced a conceptual framework as part of an on-going work for developing a framework for supporting sampling activities in SE surveys. These concepts were tailored based on lessons learned in undertaking SE surveys, applying systematic strategies for retrieving and recruiting subjects through professional social networks [6,7]. We also briefly presented an analysis of web-based sources of sampling in the light of the presented concepts. As a next step, we intend to develop the processes for supporting the referred framework and evaluate its feasibility through case studies.

5. ACKNOWLEDGMENTS

We thank CNPq for supporting our research. We also thank the SERG group from Lund University by the relevant contributions to this research.

6. REFERENCES

[1] Sjøberg, D. I. et al. 2005. A survey of controlled experiments in software engineering. *IEEE Transactions on Software Engineering*, 31(9): 733-753.

[2] Dybå, T., Kampenes, V. B. and Sjøberg, D. I. K. 2007. A systematic review of statistical power in software engineering experiments. *Inf. and Soft. Technology* 48: 745–755.

[3] Conradi, R. et al. 2005. Reflections on conducting an international survey of Software Engineering. *International Symposium on Empirical Software Engineering*, 214-223.

[4] Pickard, L. M., Kitchenham, B. A. and Jones, P. W. 1998. Combining empirical results in software Engineering. *Information and Software Technology* 40: 811–821.

[5] de Mello, R. M and Travassos, G. H. 2013. An Ecological Perspective Towards the Evolution of Quantitative Studies in Software Engineering. In: Proc. of *17th EASE*, 216-219.

[6] de Mello, R. M. and Travassos, G. H. 2013. Would Sociable Software Engineers Observe Better? In: *Proc. of 7th ESEM*, 279-282, IEEE.

[7] de Mello, R. M., da Silva, P. C. and Travassos, G. H. 2014. Investigating Probabilistic Sampling Approaches for Large-Scale Surveys in Software Engineering. In: *11th Workshop on Experimental Software Engineering (ESELAW)*.

[8] Wohlin, C. et al. 2012. Experimentation in Software Engineering. *Springer*.

[9] Frakes, W. B., and Fox, C. J. 1995. Sixteen questions about software reuse. 1995. *Communications of the ACM* 38: 75.

[10] Särndal C. A., Swensson, B., and Wretman, J. 1992. Model Assisted Survey Sampling. *Springer*.

[11] Kitchenham, B. A. et al. 2010. Systematic literature reviews in software engineering – A tertiary study. *Information and Software Technology* 52: 792–805.

[12] Stolee, K. T. and Elbaum S. 2010. Exploring the Use of Crowdsourcing to Support Empirical Studies in Software Engineering. In: *Proc. of 4th ESEM*.

[13] Stolee, K. T. and Elbaum S. 2013. On the Use of Input/Output Queries for Code Search. *In: Proc. 7th ESEM*.