Reproducibility of Survey Results: A New Method to Quantify Similarity of Human Subject Pools

Atieh R. Khamesi[†], Riccardo Musmeci[†], Simone Silvestri[†], and D. A. Baker[§]
[†]University of Kentucky, Lexington, USA, [§]Missouri University of Science and Technology, Rolla, USA
{atieh.khamesi,riccardo.musmeci,simone.silvestri}@uky.edu, and bakerden@mst.edu

***Abstract*—Smart Connected Communities (SCCs) is a novel paradigm that brings together multiple disciplines, including social-sciences, computer science, and engineering. Large-scale surveys are a fundamental tool to understand the needs and impact of new technologies to human populations, necessary to realize the SCC paradigm. However, there is a growing debate regarding the reproducibility of survey results. As an example, it has been shown that surveys may easily provide contradictory results, even if the subject populations are statistically equivalent from a demographic perspective. In this paper, we take the initial steps towards addressing the problem of reproducibility of survey results by providing formal methods to quantitatively justify apparently inconsistent results. Specifically, we define a new dissimilarity metric between two populations based on the users answers to non-demographic questions. To this purpose, we propose two algorithms based on submodular optimization and information theory, respectively, to select the most representative questions in a survey. Results show that our method effectively identifies and quantifies differences that are not evident from a purely demographic point of view.**

***Index Terms*—Reproducibility, Surveys, Dissimilarity Metrics.**

I. Introduction

Smart Connected Communities (SCCs) is a novel paradigm recently coined by the National Science Foundation (NSF) [1]. According to this paradigm, communities can exploit technological advances, such as ubiquitous connectivity, big data analysis, and sensing technologies, to improve well-being and prosperity [2]. The SCCs paradigm clearly has an interdisciplinary nature, bringing together multiple disciplines including social-sciences, computer science, economics, and engineering.

A fundamental challenge in SCCs is the understanding of the complex interaction between technology and society [1]. In fact, such *social dimension* is often neglected in engineering and technological works [3]. Nevertheless, the success of novel technologies is strictly related to the understanding of social factors such as responses to and acceptance of technological advances, behavioral changes due to availability of new information, as well as short and long term impact of new technologies on communities [4], [5].

In order to study such social aspects, large-scale surveys have been widely used as a scientific tool. However, there is a growing debate regarding the *reproducibility* of survey results in particular, and scientific results in general [6], [7]. In fact, reproducibility is at the core of the scientific method [8] and it refers to the possibility of repeating an experiment independently, in order to corroborate or confute the scientific findings.

A distinguished survey [9], in which over 1500 scientists answered questions about the importance of reproducibility, showed that 90% of the respondents believe that there is a replication crisis at hand. This problem is especially important in social science fields, where scientists aim to quantify and understand complex human behaviors in hopes of making reliable predictions about future behaviors. A 2015 study conducted by Open Science Collaboration looked at 100 psychology studies from 2008 and claimed they were only able to replicate about 40% of the findings [10].

There is a swell of interest in increasing rates and likelihood of reproducibility with suggested approaches ranging from efforts to increase publication value for replication studies and null finding [11], to changing interpretation of p-values [12], [13], to pre-registration [14].

What has not been explored is how to quantitatively characterize similarities and dissimilarities between the original subject pool and the subject pool being used for the replication study. Generally, to make a broad determination of sample similarity, *demographic* information is used, such as gender, ethnicity, socio-economic status, and so on. However, such information is often not sufficient in explaining dissimilarities between subject pools.

In this paper we go *beyond* demographics qualities and propose a *dissimilarity metric* to compare two potentially different populations. This measure may help explain incongruousness and inconsistencies of survey replication results, by showing the "dissimilarity" of two populations that are potentially equivalent from a demographic perspective.

The scenario we envision is the following. A survey is completed and its results are public (questions and anonymous answers). When a second survey is designed to further investigate or corroborate the previous findings, some questions that characterize the first population are selected from the initial survey and included in the second survey. By comparing the answers to these selected questions, we can calculate our dissimilarity metric between the populations. This way we are able to provide a quantitative justification for the potentially incoherent or contradictory results. It is worth mentioning that it is important to limit the number of selected questions. In fact, it has been shown that asking too many questions is not only cumbersome to respondents, but may even cause deterioration in research quality [15]. To the best of our

knowledge, this is the first paper that addressed the issue of reproducibility in survey results through the exploitation of algorithmic and optimization techniques, and it represents a first step towards understanding and mitigating this problem.

II. Related Work

Reproducibility of scientific results particularly in social science and psychology contexts has received significant attention in recent years. In [10], [16], [17], the authors focused on replicating selected studies from psychology journals, Nature and Science journals. To evaluate the reproducibility of such studies, they used measures such as significance, p-values, Bayesian analysis, and prediction markets. Particularly, in [10] they claim that no single indicator sufficiently describes replication success, and their indicators are not the only ways to assess reproducibility. Furthermore, the results obtained by Bayesian analysis in [16] suggest that research community could predict which results would replicate and that failures to replicate were not the result of chance alone.

A remarkable work in this area is the *Reproducibility Project* [18], an open large-scale collaborative effort to systematically examine the rate and predictors of reproducibility in psychological science. In a first attempt of estimating reproducibility of studies published in top psychology journals, it was found that only 39% of them could be unambiguously reproduced. This project has significantly captured the attention of the scientific community [19], [20]. For example, in [19] they used Bayesian analysis to conclude that the apparent failure of the Reproducibility Project to replicate many target effects can be adequately explained by overestimation of effect sizes due to small sample sizes and publication bias in the psychological literature. Furthermore, in [20] the obtained results showed that researchers, replicators, and consumers should be mindful of contextual factors that might affect a psychological process in conducting a replicable study.

Note that, there exist some works studying the selection of participants in both intercept and online surveys [21], [22], [23]. However, their main goal is to avoid selection bias and they do not consider the reproducibility of survey results.

In this paper, we present a different approach to provide an additional tool towards the replication and reproducibility of large-scale surveys. Specifically, we provide optimization and algorithmic solutions to identify the most relevant questions in a survey, which are used to calculate a dissimilarity metric between population pools that goes beyond standard statistical demographic sampling. Note that, some surveys include attention and manipulation checks, where some questions are repeated, at different stages, to differentiate between participants who paid attention, and those who did not. As such, these types of questions would not likely make good questions to calculate a dissimilarity metric, as they would have very little variability among participants.

III. Survey Methodology

In social science, survey question sets are widely used as a tool to investigate human behaviors both in quantitative and qualitative research [24], [25]. Some common forms of questions include: (i) *Multiple choice - categorical questions* that ask respondents to select an answer(s) from a pre-defined set of categorical or nominal items. (ii) *Multiple choice - continuous questions* that ask respondents to select an answer that fits along a continuum or ordered range such as choices that move from least to greatest. Researchers consider when and how to employ these approaches based on consideration of existing literature that informs their domain and particular research question, typically with goals such as reducing bias and increasing reliability [25], [24].

In this paper we focus on these two forms of discrete choice questions. We will analyze additional questions types such as ranked responses in future work.

IV. Survey Model and Problem Formulation

In this work, a survey is defined as a tuple, $S = (Q, U)$, where $Q = \{q_1, q_2, \ldots, q_k\}$ is the set of questions, and $U = \{u_1, u_2, \ldots, u_n\}$ represents the subject pool, referred to "users" in the following. To calculate our dissimilarity metric, the first step is to identify the questions that best characterize a population participating in a survey, with respect to their answers. We model this through the concept of partition implied by a set questions $Q' = \{q'_1, \ldots, q'_l\} \subseteq Q$.

Definition 1 (Partition of a set)**.** *Given a survey S and a set of questions $Q' = \{q'_1, \ldots, q'_l\} \subseteq Q$, Q' implies a partition $\mathcal{C}(Q') = \{C_1, C_2, \ldots\}$ defined as the set of nonempty disjoint* classes *of users such that every user $u \in U$ is in exactly one of these classes, and users in the same class gave the same answers to all the questions in Q'.*

According to our model, a population is characterized by a partition, also refer to as class, with respect to their answers. Any subset of Q gives a partition of user with likely different classes. Note that, since users in a specific class gave the same answers to all questions, they are *indistinguishable*. Intuitively, we want to minimize this ambiguity when selecting the most informative questions. To this purpose, given a partition $\mathcal{C}(Q')$, we measure its quality by the largest class $C^*_{(Q')}$ defined as

$$C^*_{(Q')} = \max_{C_i \in \mathcal{C}(Q')} |C_i|. \tag{1}$$

Clearly, considering the entire set of questions, Q, gives the best possible partition. However, when designing the second survey, this trivial solution is not practical for the following reasons. First, research quality can be deteriorated when participants are asked to respond to too many questions in one setting [15], [26]. As such adding this subset of questions should not unnecessarily increase the length of the follow-up. An additional practical consideration is that, in cases where participants compensation is determined by the time it takes to complete the survey, superfluous question have a financial cost as well. Finally, online survey platforms such as Qualtrics [27] charge an amount of money proportional to the number of question in a survey.

Taking into account the above considerations, our first problem is to find the best subset $Q^* \subseteq Q$, that minimizes

the uncertainty in discerning individuals in U, under a budget constraint for the number of questions that can be selected. Formally, we look for the set $Q^* \subseteq Q$ with minimum $|C^*_{(Q^*)}|$, within a budget B, that is:

Problem 1. *Given a survey $S = (Q, U)$ and a budget B, the best partition is given by*

$$Q^* = \min_{Q' \subseteq Q} |C^*_{(Q')}| \qquad s.\ t. \quad |Q'| \leq B. \tag{2}$$

Example 1. *In the following, we provide a toy example to clarify the problem formulation and the impact of different subset of questions in the partition of a set of users U. Consider a survey with 8 users and 4 questions, where each question has exactly 6 possible answers (numbered from 0 to 5, for simplicity). In Table I, we exemplify the users' answers to each question.*

Let us consider a set $Q' = \{q_1\}$. This would result in a partition set $\mathcal{C}(\{q_1\})$ given by:

$$\mathcal{C}(\{q_1\}) = \{\{u_1, u_2, u_7\}, \{u_3, u_6\}, \{u_4, u_8\}, \{u_5\}\}. \tag{3}$$

The largest class in this case is $\{u_1, u_2, u_7\}$, which is implied by the fact that all these three users gave answer 5 to question q_1. By adding q_2 to $Q' = \{q_1\}$, we obtain the following partition:

$$\mathcal{C}(\{q_1, q_2\}) = \{\{u_1, u_7\}, \{u_2\}, \{u_3, u_6\}, \{u_4\}, \{u_5\}, \{u_8\}\}. \tag{4}$$

The size of the largest class has reduced to two users, since adding q_2 resolved some ambiguity. However, not all questions contribute in the reduction of the size of the largest class. In fact, by including q_3 we obtain the same partition as in Eq. (4). If we assume a budget $B = 3$, the optimal solution in this example is $Q^* = \{q_1, q_2, q_4\}$. In fact, in this case, every user can be uniquely identified by their answers,

$$\mathcal{C}(\{q_1, q_2, q_4\}) = \{\{u_1\}, \{u_2\}, \{u_3\}, \{u_4\}, \{u_5\}, \{u_6\}, \{u_7\}, \{u_8\}\}. \tag{5}$$

TABLE I
SURVEY EXAMPLE

user id	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8
q_1	5	5	3	2	4	3	5	2
q_2	3	2	1	4	3	1	3	1
q_3	1	3	0	1	3	0	1	0
q_4	0	0	1	2	0	0	2	2

In the next sections, we propose two algorithms to solve the optimization problem in (2).

V. ALGORITHMS FOR QUESTION SELECTION

In this section we discuss the two algorithms used to select the most informative questions from a survey.

A. The MIMOSA Algorithm

The first algorithm is named MIn Max SubmOdulAr *(MIMOSA)* algorithm. In order to define MIMOSA, we reformulate the problem in Eq. (2) into an equivalent maximization problem. To this purpose, we rewrite the objective function as $|U| - |C^*_{(Q')}|$ and thus the problem becomes:

$$Q^* = \max_{Q' \subset Q} |U| - |C^*_{(Q')}| \qquad \text{s. t.} \quad |Q'| \leq B \tag{6}$$

The following corollary shows that the two problems have the same optimal solution.

Corollary 1. *The problems defined in Eqs.* (2) *and* (6) *have the same optimal solution.*

Given the equivalency between the two problems, we now focus on the maximization problem in Eq. (6) and show that its objective function is *submodular*. A function is *submodular* according to this definition.

Definition 2 (Submodular function [28])**.** *Given a finite ground set Q and a function $f : 2^Q \rightarrow R_+$, then $\forall\ Q_1 \subseteq Q_2 \subseteq Q, q \in Q \setminus Q_2, f$ is submodular iff*

$$f(Q_1 + q) - f(Q_1) \geq f(Q_2 + q) - f(Q_2),$$

where the notation $Q_i + q$ stands for $Q_i \cup \{q\}$.

The following theorem shows that the objective function of the maximization problem in Eq. (6) is submodular, under the assumption that users answer questions independently. Note that, this assumption is only considered for the proof and the MIMOSA algorithm does not rely on the independent answers. However, the proof for the general case, considering potentially dependent answers, is strongly supported by experimental evidence, but it is still an open problem.

Theorem 1. *Let Q and U be the sets of questions and users, respectively. Given the fact that the users answer to each question independently, the objective function of the problem in Eq.* (6) *is* submodular.

Proof: To prove the Theorem we need to show that, given two sets of questions Q_1 and Q_2 such that $Q_1 \subseteq Q_2 \subseteq Q$ and $q \in Q \setminus Q_2$, it holds the following:

$$|U| - |C^*_{(Q_1+q)}| - (|U| - |C^*_{(Q_1)}|) \geq |U| - |C^*_{(Q_2+q)}| - (|U| - |C^*_{(Q_2)}|),$$

This equation can be rewritten as:

$$|C^*_{(Q_1)}| - |C^*_{(Q_1+q)}| \geq |C^*_{(Q_2)}| - |C^*_{(Q_2+q)}|. \tag{7}$$

Using Eq. (1) we have,

$$\max_{C_i \in \mathcal{C}(Q_1)} |C_i| - \max_{C_i \in \mathcal{C}(Q_1+q)} |C_i| \geq \max_{C_i \in \mathcal{C}(Q_2)} |C_i| - \max_{C_i \in \mathcal{C}(Q_2+q)} |C_i|. \tag{8}$$

Let us define P^q_j as the probability of a user to answer j to a question q. Let C_j be the class of users that answered j, this probability can be expressed as $P^q_j = |C_j|/|U|$. We extend

this notation to a generic set of questions Q_i, by defining $P_j^{Q_i}$ as the probability of providing the set of answers j to the questions in Q_i. Moreover, since we assumed that users answer questions independently, for any additional questions $q \in Q \setminus Q'$, we have $P_k^{Q_i+q} = P_j^{Q_i} \cdot P_l^q$, where the set of answers k combines the answers j and l to the questions in Q_i and q, respectively. Accordingly, by dividing both sides of Eq. (8) by $|U|$ we obtain:

$$\max_i\{P_i^{Q_1}\} - \max_{i,j}\{P_i^{Q_1} \cdot P_j^q\} \geq \max_i\{P_i^{Q_2}\} - \max_{i,j}\{P_i^{Q_2} \cdot P_j^q\}. \quad (9)$$

Without loss of generality, we assume that answers are numbered so that the first answer provides the largest class, and therefore the largest probability, that is: $P_1^{Q_1} \geq P_i^{Q_1},\ \forall i \in \{2,\dots,n_{Q_1}\}, P_1^{Q_2} \geq P_i^{Q_2},\ \forall i \in \{2,\dots,n_{Q_2}\}, P_1^q \geq P_i^q,\ \forall i \in \{2,\dots,n_q\}$ where n_{Q_i} is number of the possible answers to questions in Q_i. Therefore, it follows

$$P_1^{Q_1} - P_1^{Q_1} \cdot P_1^q \geq P_1^{Q_2} - P_1^{Q_2} \cdot P_1^q$$
$$P_1^{Q_1}(1 - P_1^q) \geq P_1^{Q_2}(1 - P_1^q) \ \Rightarrow\ P_1^{Q_1} \geq P_1^{Q_2},$$

which is always true since $Q_1 \subseteq Q_2$. Therefore, the objective function of the problem is submodular. ∎

Following recent advancements in submodular optimization theory [29] we can define the MIMOSA algorithm as a greedy algorithm as shown in Algorithm 1. MIMOSA builds the solution Q_M iteratively, and initially $Q_M = \emptyset$ (line 2). At each iteration, the algorithm selects the question $q \in Q \setminus Q_M$ that maximizes the gain in the objective function (line 4). The algorithm terminates when $|C^*(Q_M)| = 1$ or the budget is exhausted (line 3).

Algorithm 1 The MIMOSA algorithm

Input: $S = (Q, U)$, B
Output: $Q_M \subset Q$

1: **procedure** MIMOSA(S, B)
2: $Q_M \leftarrow \emptyset$
3: **while** $|C^*_{(Q_M)}| > 1$ AND $|Q_M| \leq B$ **do**
4: $q_M \leftarrow \arg\max_{q \in Q \setminus Q_M}\{|C^*_{(Q_M)}| - |C^*_{(Q_M+q)}|\}$
5: $Q_M \leftarrow Q_M + q_M$
6: **return** Q_M

The submodularity of the objective function allows us to prove that MIMOSA provides a $1 - \frac{1}{\sqrt{e}}$ approximation bound for the problem in Eq. (6).

Theorem 2 ([29]). *Given a finite ground set Q, a submodular function $f : 2^Q \rightarrow R_+$, and a budget B. If f is non-decreasing and $f(\emptyset) = 0$, then the greedy algorithm produces a solution which is at least $1 - \frac{1}{\sqrt{e}}$ times the optimal value.*

We prove that the objective function of the maximization problem in (6) satisfies the conditions of Theorem 2.

Lemma 1. *The objective function of the problem in Eq.* (6) *is submodular, non-decreasing and it holds that $|U| - |C^*_{(\emptyset)}| = 0$.*

Proof: Theorem 1 proves that the objective function is submodular. In order to show that it is non-decreasing, let us add a question $q \in Q \setminus Q'$ to an arbitrary set Q'. The classes of users identified by the partition implied by $Q' + q$ are either the same as Q' (like the example in Section IV when adding q_3 to $\{q_1, q_2\}$) , or otherwise some classes may be further partitioned into smaller sets. Since this holds for all classes, it also holds for the largest class, i.e. $C^*_{(Q'+q)} \leq C^*_{(Q')}$. Therefore, $|U| - |C^*_{(Q')}| \leq |U| - |C^*_{(Q'+q)}|, \forall Q',\ q \in Q \setminus Q'$, which proves that the function is non decreasing..

To conclude the proof of the Lemma, we point out that for an empty set of questions $Q' = \emptyset$, the largest class corresponds to the entire set U, i.e., $C^*_{(\emptyset)} = U$. Therefore, the objective function is $|U| - |C^*_{(\emptyset)}| = 0$. ∎

B. The JINGO Algorithm

In this section, we describe the JoInt eNtropy alGOrithm *(JINGO)*. JINGO adopts the concept of *joint entropy* from Information Theory, which is defined as follows.

Definition 3 (Joint Entropy [30])**.** *Given a set of discrete random variables, $X_1, \dots, X_n$ the joint entropy is given by*

$$H(X_1, \dots, X_n) = \sum_{x_1,\dots,x_n} \mathbf{P}(x_1, \dots, x_n) \frac{1}{\log_2(\mathbf{P}(x_1, \dots, x_n))},$$

where $\mathbf{P}(x_1, \dots, x_n)$ is the probability of $X_i = x_i$, for $i = 1, \dots, n$.

The joint entropy measures the uncertainty of random variables, and it is maximum for the uniform distribution [30]. In our case, we can interpret the questions in Q as discrete random variables, where the possible answers represent the observed values of such variables. Given a subset of questions $Q' \subseteq Q$, the partition $\mathcal{C}(Q') = \{C_1, \dots, C_l\}$ can be seen as a probability distribution over the set of all possible answers to the questions in Q'. Since $\mathcal{C}(Q')$ is a partition of U, the probability value of each class can be obtained by dividing the size of that class by U.

Algorithm 2 The JINGO Algorithm

Input: $S = (Q, U)$, B
Output: $Q_J \subset Q$

1: **procedure** JINGO(S, B)
2: $Q_J \leftarrow \emptyset$
3: **while** $|C^*_{(Q_J)}| > 1$ AND $|Q_J| \leq B$ **do**
4: $q_J \leftarrow \arg\max_{q \in Q \setminus Q_J}\{\mathrm{H}(Q_J + q)\}$
5: $Q_J \leftarrow Q_J + q_J$
6: **return** Q_J

The joint entropy of a set Q' is higher as the distribution is more uniform, i.e. users are equally distributed in the classes. As a result, the idea of JINGO is to select questions in order to maximize the joint entropy. This also implies a reduction

in the size of the largest class, which is the objective of the optimization problem in Eq. (6).

The pseudocode of the algorithm is shown in Algorithm 2. JINGO iteratively builds the solution Q_J by selecting the question, $q \in Q \setminus Q_J$, that provides the maximum increase in the joint entropy value (line 4). Note that, JINGO and MIMOSA can provide, in general, different solutions. In fact, the joint entropy considers the entire distribution, while MIMOSA only focuses on the size of the largest class.

C. A Note on Complexity

The pseudo codes of MIMOSA and JINGO show similarities between the structure of the two algorithms, which are reflected in a similar computational complexity. Specifically, both while loops can iterate at most B times. Finding the question that maximizes the reduction of the largest class (line 4, MIMOSA), or the question that provides maximum joint entropy (line 4, JINGO), have the same complexity. In fact, they both require a time proportional to the number of questions left and the number of classes. While the number of questions left is upper bounded by Q, the number of classes may potentially grow exponentially at each iteration. Nevertheless, we point out that the actual number of non-empty classes can be at most $|U|$, since classes are disjoint and users can be at most in one class. As a result, the complexity of MIMOSA and JINGO can be expressed as $O(B \times |Q| \times |U|)$.

VI. A New Dissimilarity Metric

In this section, we discuss the mathematical formulation of our dissimilarity metric between two subject pools. We recall that, according to our approach, the solution of the optimization problem in Eq. (6) returns a subset of questions Q^* from the first survey (MIMOSA or JINGO could be used in practice for an approximated solution). When a second survey is designed, the questions in Q^* are also included. The dissimilarity metric quantifies the difference between the answers to the common questions once the second survey is completed. The metric is defined through a weighted bipartite graph $G = (L, R, E)$. The nodes in L and R represent the users in the first and second survey, respectively. There is an edge $e \in E$ for each pair of nodes (l, r), such that $l \in L$ and $r \in R$. The edge is weighted according to the dissimilarity between the answers provided to the questions in Q^*.

Intuitively, this dissimilarity must take into account the different types of questions. We consider two types of questions in this work, namely *Multiple choice - categorical questions*, that ask users to select an answer from a pre-defined set of categorical or nominal items, and *Multiple choice - continuous questions*, that ask users to select an answer that fits along a continuum or ordered range such as choices that move from greatest to least. Let us consider two users l and r, and a categorical questions $q_i \in Q^*$ which was answered l_i and r_i, respectively. Answers to categorical questions usually do not follow an order, such as for example questions regarding gender or ethnicity. As a result, we consider two answers to be either the same ($l_i = r_i$), contributing zero to the dissimilarity metric, or different ($l_i \neq r_i$), contributing one to the dissimilarity metric. On the contrary, continuous questions present a natural order. As an example, one of such questions may ask to rate the level of agreement with a statement using a Likert scale (strongly agree / agree / don't know / disagree / strongly disagree). In this case, we assign a numerical value to each possible answer following this order. Let us consider the same users, and a continuous question $q_j \in Q^*$, which was answered l_j and r_j. The contribution to the dissimilarity metric is given by the normalized distance between the answers, that is $(|l_j - r_j|)/||q_j||$, where $||q_j||$ is the range of the numeric interval assigned to the answers.

Let $I_{cat} \subseteq Q^*$ be the subset of categorical questions and $I_{cont} \subseteq Q^*$ be the subset of continuous questions, the dissimilarity of two users $l \in L$ and $r \in R$ is defined as

$$d(l, r) = \sum_{q_i \in I_{cat}} (l_i \neq r_i) + \sum_{q_j \in I_{cont}} \frac{|l_i - r_i|}{||q_j||}. \quad (10)$$

This value is used to weight the edges in the bipartite graph. To define the dissimilarity metric between the subject pools of the first and second survey, we want match similar users, and then measure the overall difference. Therefore, we define the dissimilarity metric between survey pools as the value of the maximum bipartite matching of the graph G. This can be found by means of the *Hungarian* algorithm [31].

VII. Evaluation

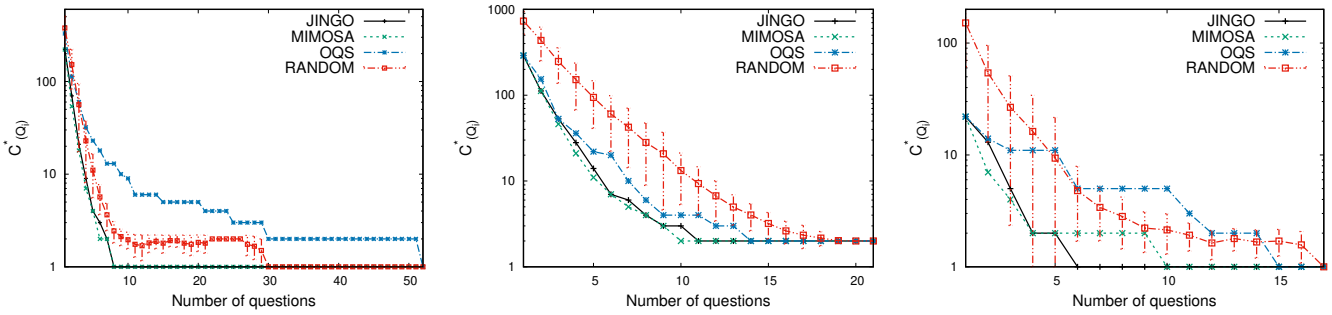
We use three real surveys on different topics. The first survey focuses on attitudes towards mental health and frequency of mental health disorders in the tech workplace [32]. The second survey explores the preferences, interests, habits, opinions, and fears of young people [33]. Finally, the last survey is a research study on user perception about common household appliances [34]. Table II summarizes the information of each survey. We consider the heterogeneity of the topics covered by these surveys a good test supporting the methodology proposed in this paper. The interested reader is referred to the cited references for more details.

TABLE II
Summary of the surveys

Survey	Mental Health in Tech [32]	Young People [33]	Household appliances [34]
Questions	27	150	17
Avg. Nr Answers.	10.44	6.63	25
Std. Dev.	16.88	6.41	31.96
Nr. of Users	1259	1010	357

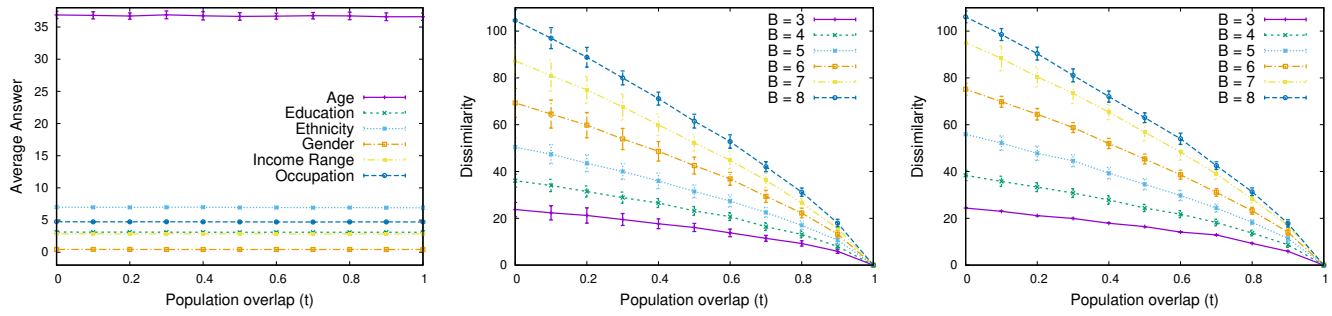
A. Questions Selection

To evaluate the performance of MIMOSA and JINGO, we compare them with two other algorithms. The first algorithm is based on the intuition that questions with a higher number of possible answers are more likely to partition users in smaller groups. As a result, this algorithm sorts questions in decreasing order of the number of possible answers. Then, it adds questions to the solution following that order, until



(a) Young people survey (b) Mental health survey (c) Household appliances survey

Fig. 1. Questions selection results: size of the largest class versus the available budget.



(a) Demographic Information (b) MIMOSA (c) JINGO

Fig. 2. Dissimilarity: (a) Demographic statistics of $Z(t)$, and Dissimilarity metric for (b) MIMOSA and (c) JINGO.

the budget is met or the size of the largest class is one. We refer to this approach as Ordered-base Questions Selection (*OQS*). The second algorithm is a random baseline approach. The algorithm simply randomly selects questions until the same termination condition of OQS are met. We refer to this algorithm as *RANDOM*.

As observed in Figs. 1 (a)-(c), RANDOM and OQS struggle in identifying the most informative questions for all the considered surveys. The fact that a random approach performs poorly may not be surprising, nevertheless it shows that a smart optimization can significantly improve the performance. As a numerical example, in the Young people survey shown in Fig. 1 (a), MIMOSA and JINGO only need 8 questions to uniquely identify all users, while RANDOM needs 30. Interestingly, OQS performs worse than random in some circumstances, as for the Young people survey and for some values of the budget in the Household appliances survey (Fig. 1 (c)). This result reveals that the actual distribution of answers is not necessarily affected by the number of possible answers, and it is of primary importance to identify the most important questions by looking at the actual distribution. In all the considered scenarios, MIMOSA and JINGO significantly outperform the other algorithms and achieve desirable performance by requiring only few questions to uniquely identify a user. For example, in Fig. 1 (a) both algorithms only need 8 questions out of 150, which is less than 6% of the whole set of questions in the survey. As the results show, there is no clear winner between MIMOSA and JINGO, since the performance are slightly affected by the distribution of the answers, which is specific of each survey.

B. Dissimilarity Metric

In the following, we show results to support the ability of our dissimilarity metric to capture differences in subject pools which are not evident from demographic statistics. Due to space scarcity and similarity of the results, we show the results for the Household appliances survey.

The experiments are performed as follows. We simulated a scenario of two surveys by splitting the original population U in two halves X and Y. Given these two disjoint sets and a number $t \in [0, 1]$, we create a new set of users $Z(t)$ by picking a fraction of t users from X, and $1 - t$ users from Y. We use MIMOSA and JINGO to identify the most representative questions under a budget constraint considering the users in X. Then, we calculate the dissimilarity metric using the bipartite graph approach discussed in Section VI between X and $Z(t)$, for a given value of t. Intuitively, t represent the fraction of overlap between X and $Z(t)$. When $t = 0$, X and Z are disjoint ($X \cap Z(0) = \emptyset$), thus the dissimilarity is expected to be maximum. Conversely, when $t = 1$, $X = Z(1)$, thus the dissimilarity is expected to be zero (i.e., each user is matched with him/herself). We averaged the results over multiple runs, since users are assigned randomly to $Z(t)$.

Fig. 2 (a) shows the average and standard deviation of

several demographic information of the users in $Z(t)$, under different values of t. Note that, in order to provide a quantifiable average for such information, we assigned a numerical value to non-numerical answers (e.g., gender and ethnicity). In these cases, first we sort the answers alphabetically and then assign a non-negative integer value to each possible answer. Conversely, Figs. 2 (b) and (c) show the value of the dissimilarity metric calculated between X and $Z(t)$ under different budget values, i.e., number of selected questions, for MIMOSA and JINGO, respectively. As expected, the metric decreases as the fraction of overlap t increases.

VIII. Conclusion

In this paper, we take the initial steps towards addressing the problem of reproducibility of survey results by providing formal methods to quantitatively justify apparently inconsistent or even contradictory results. Specifically, we define a new dissimilarity metric between two populations, based on the users' answers to non-demographic questions. To this purpose, we propose two algorithms named MIMOSA and JINGO, which are based on submodular optimization and information theory, respectively. The selected questions can be included in other surveys answered by a potentially different population. Results show that our method effectively identifies and quantifies differences that are not evident from a demographic point of view.

Acknowledgment

This work is supported by the National Institute for Food and Agriculture (NIFA) under the grant 2017-67008-26145, the NSF grant EPCN-1936131, and the NSF CAREER grant CPS-1943035.

References

[1] S. H. Plimpton, "The smart and connected communities (scc) program solicitation," 2019. [Online]. Available: https://www.nsf.gov/pubs/2019/nsf19564/nsf19564.htm

[2] V. K. Shah, S. Bhattacharjee, S. Silvestri, and S. K. Das, "Designing sustainable smart connected communities using dynamic spectrum access via band selection," in *Proceedings of the 4th ACM International Conference on Systems for Energy-Efficient Built Environments*, 2017, pp. 1–10.

[3] A. R. Khamesi, S. Silvestri, D. Baker, and A. D. Paola, "Perceived-value-driven optimization of energy consumption in smart homes," *ACM Transactions on Internet of Things*, vol. 1, no. 2, pp. 1–26, 2020.

[4] A. R. Khamesi, E. Shin, and S. Silvestri, "Machine learning in the wild: The case of user-centered learning in cyber physical systems," in *2020 International Conference on COMmunication Systems & NETworkS (COMSNETS)*. IEEE, 2020, pp. 275–281.

[5] E. Shin, A. R. Khamesi, Z. Bahr, S. Silvestri, and D. A. Baker, "A user-centered active learning approach for appliance recognition," in *2020 International Conference on COMmunication Systems & NETworkS (COMSNETS)*. IEEE, 2020.

[6] F. L. Cook, "Dear colleague letter: Achieving new insights through replicability and reproducibility," 2018. [Online]. Available: https://www.nsf.gov/pubs/2018/nsf18053/nsf18053.jsp

[7] IEEE Workshop on The Future of Research Curation and Research Reproducibility. [Online]. Available: https://www.ieee.org/publications/research-reproducibility.html

[8] M. McNutt, "Reproducibility," *Science*, vol. 343, no. 6168, pp. 229–229, 2014. [Online]. Available: http://science.sciencemag.org/content/343/6168/229

[9] M. Baker, "1,500 scientists lift the lid on reproducibility," *Nature News*, vol. 533, no. 7604, p. 452, 2016.

[10] Open Science Collaboration, "Estimating the reproducibility of psychological science," *Science*, vol. 349, no. 6251, p. 4716, 2015.

[11] V. E. Johnson, R. D. Payne, T. Wang, A. Asher, and S. Mandal, "On the reproducibility of psychological science," *Journal of the American Statistical Association*, vol. 112, no. 517, pp. 1–10, 2017.

[12] B. B. McShane, D. Gal, A. Gelman, C. Robert, and J. L. Tackett, "Abandon statistical significance," *The American Statistician*, vol. 73, no. sup1, pp. 235–245, 2019.

[13] D. J. Benjamin, J. O. Berger *et al.*, "Redefine statistical significance," *Nature Human Behaviour*, vol. 2, 09 2017.

[14] B. A. Nosek, C. R. Ebersole, A. C. DeHaven, and D. T. Mellor, "The preregistration revolution," *Proceedings of the National Academy of Sciences*, vol. 115, no. 11, pp. 2600–2606, 2018.

[15] M. Galesic and M. Bosnjak, "Effects of questionnaire length on participation and indicators of response quality in a web survey," *Public opinion quarterly*, vol. 73, no. 2, pp. 349–360, 2009.

[16] C. Camerer, A. Dreber, F. Holzmeister, T. Ho, J. Huber, M. Johannesson, M. Kirchler, G. Nave, B. Nosek, T. Pfeiffer, A. Altmejd, N. Buttrick, T. Chan, Y. Chen, E. Forsell, A. Gampa, E. Heikensten, L. Hummer, T. Imai, and H. Wu, "Evaluating the replicability of social science experiments in nature and science between 2010 and 2015," *Nature Human Behaviour*, vol. 2, 08 2018.

[17] A. Dreber, T. Pfeiffer, J. Almenberg, S. Isaksson, B. Wilson, Y. Chen, B. A. Nosek, and M. Johannesson, "Using prediction markets to estimate the reproducibility of scientific research," *Proceedings of the National Academy of Sciences*, vol. 112, no. 50, pp. 15 343–15 347, 2015.

[18] Reproducibility project. [Online]. Available: https://osf.io/ezcuj/wiki/home/

[19] A. Etz and J. Vandekerckhove, "A bayesian perspective on the reproducibility project: Psychology," *PLOS ONE*, vol. 11, no. 2, pp. 1–12, 02 2016. [Online]. Available: https://doi.org/10.1371/journal.pone.0149794

[20] J. J. Van Bavel, P. Mende-Siedlecki, W. Brady, and D. Reinero, "Contextual sensitivity in scientific reproducibility," *Proceedings of the National Academy of Sciences*, vol. 113, p. 201521897, 05 2016.

[21] Z. A. Greenacre *et al.*, "The importance of selection bias in internet surveys," *Open Journal of Statistics*, vol. 6, no. 03, p. 397, 2016.

[22] M. P. Battaglia, D. A. Dillman, M. R. Frankel, R. Harter, T. D. Buskirk, C. B. McPhee, J. M. DeMatteis, and T. Yancey, "Sampling, data collection, and weighting procedures for address-based sample surveys," *Journal of Survey Statistics and Methodology*, vol. 4, no. 4, pp. 476–500, 2016.

[23] K. J. Papacostas and J. Foster, "A replication approach to controlled selection for catch sampling intercept surveys," *Fisheries Research*, vol. 229, p. 105609, 2020.

[24] P. S. Visser, J. A. Krosnick, and P. J. Lavrakas, "Survey research." 2000.

[25] J. A. Krosnick, "Survey research," *Annual review of psychology*, vol. 50, no. 1, pp. 537–567, 1999.

[26] A. R. Herzog and J. G. Bachman, "Effects of questionnaire length on response quality," *Public opinion quarterly*, vol. 45, no. 4, pp. 549–559, 1981.

[27] "Qualtrics," https://www.qualtrics.com.

[28] A. Krause and D. Golovin, "Submodular function maximization." 2014.

[29] A. Krause and C. Guestrin, "A note on the budgeted maximization of submodular functions," *Tech. Report at CMU*, 2005.

[30] T. M. Cover and J. A. Thomas, *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. New York, NY, USA: Wiley-Interscience, 2006.

[31] S. Skiena, *The Algorithm Design Manual*, 1997.

[32] (2014) Mental health in tech survey. [Online]. Available: https://www.kaggle.com/osmi/mental-health-in-tech-survey/home

[33] (2013) Young people survey. [Online]. Available: https://www.kaggle.com/miroslavsabo/young-people-survey

[34] (2017) Household appliances missouri science and technology. [Online]. Available: http://mst.qualtrics.com/jfe/form/SV_9nKhtCbEZPsCoMl