Automated Name Selection for the Network Scale-up Method

Adrià Fenoy*[1], Michał Bojanowski[1,2], and Miranda J. Lubbers[1]

[1]COALESCE Lab, Department of Social and Cultural Anthropology, Autonomous University of Barcelona
[2]Chair of Quantitative Methods and Information Technology, Kozminski University

February 11, 2023

Abstract

The distribution of the number of acquaintances among members of a society is a relevant feature of its social structure. Furthermore, the number of acquaintances (or "degree") is used for estimating other societal features, such as the size of hard-to-count subpopulations or social cohesion. To estimate the degree, the Network Scale-Up Method (NSUM) asks survey respondents about the number of people they know with a set of first names for which name statistics are available. For this method to be precise, a set of names needs to be selected for the survey that jointly represent the population on a smaller scale in terms of relevant traits such as gender or age. Finding the optimal set of names is a combinatorial problem for which this paper provides a solution approach. The approach can serve other NSUM users, and can be applied to any population for which name statistics distributed over different categories are available. We empirically show that our approach successfully provides subsets of names replicating the population distribution for six countries with very different name statistics.

Keywords: social networks; acquaintanceship volume; Network Scale-Up Method; combinatorial optimization

1 Introduction

How many people does a person know? The distribution of the size of individuals' acquaintanceship networks (i.e., "degree") is an essential feature of a population's social structure (e.g., Lubbers et al., 2019). Furthermore, knowledge of individuals' degrees is often needed to estimate other relevant population features, such as the size of a subpopulation for which we do not have statistics, or social cohesion (DiPrete et al., 2011).

So, how can the degree be estimated reliably? Bernard et al. (Bernard et al., 2010; Bernard et al., 1991; McCarty et al., 2001) developed the so-called Network Scale-Up Method (NSUM) to

*Email: adria.fenoy@uab.cat

estimate the size of hard-to-count or hidden populations, such as the number of people who died in an earthquake. To apply this method, they first needed to estimate the size of individual acquaintanceship networks. Assuming that people mix randomly in society, they argued that acquaintanceship networks would reflect, on a small scale, the composition of the overall population, that is, $\frac{y_{ik}}{a_i} = \frac{N_k}{n}$, where y_{ik} represents the number of people individual i knows in subgroup k, a_i the degree of individual i, N_k the size of subgroup k in the population, and N the size of the total population. Thus, by asking an individual i how many people (s)he knows in a subgroup k of known size (y_{ik}), we can estimate his/her degree a_i by dividing his/her response by the proportion of the respective subgroups in the entire population (i.e., $\frac{N_k}{n}$). To obtain a more reliable estimate, the NSUM authors suggested asking about a variety of subgroups and averaging the estimates. The method has been adopted in many empirical studies, particularly in the area of public health (cf. Bernard et al., 2010; Feehan and Salganik, 2016). Furthermore, methodological extensions have been proposed (e.g., Maltiel et al., 2015; McCormick et al., 2010).

Nonetheless, this estimation of acquaintanceship network size depends on three critical assumptions (e.g., Bernard et al., 1991; McCormick et al., 2010). First, individuals should have full knowledge about whom in their network belongs to the subgroups they are asked about. *Transmission* error occurs when a respondent might know people from a subgroup but does not know they belong to that subgroup, for instance, due to stigma or simply not knowing those people well enough. To address this error, McCormick et al. proposed asking for first name subgroups (e.g., how many people do you know whose name is Victoria?) if name statistics are available nationally, since it can be reasonably assumed that respondents know the first names of the people they know. The second assumption is that people mix randomly in society; thus, all individuals have the same probability of knowing people from a particular subgroup. *Barrier effects* appear when this is not the case. This assumption is particularly problematic because networks tend to segregate along variables such as age, gender, race, and social class (e.g., DiPrete et al., 2011), and to cluster locally (e.g., Rivera et al., 2010). By looking at the problem from a statistical perspective, McCormick et al. noted that barrier effects could be reduced by asking for subgroups whose joint distribution over relevant traits such as age and gender represents the population at a smaller scale. In the case of names, names strongly associated with particular social classes or other traits that segregate the population should be avoided. The third assumption is that respondents can reliably recall the number of people they know in a certain subgroup.

Recall error is higher when the subgroups we ask about are larger, as respondents will be more likely to under-recall the members they know. This error can thus be minimized by asking about relatively small subgroups. Thus, McCormick et al.'s directives for estimating acquaintanceship volume can be summarized as follows:

1. use first names as population subgroups when name statistics are available for a population,
2. use name subgroups whose joint distribution over traits such as gender or age is similar to a "scaled-down" population distribution across these traits, and
3. use name subgroups with a prevalence in the range $[0.1\%, 0.2\%]$ of the total population, which provides a good trade-off between recall error and the number of names needed to achieve precise estimates.

Many countries collect first-name statistics of the resident population, often broken down by gender, age group, and/or region. Directives 1 and 3 are, therefore, easily accomplished. Instead, directive 2 leads to the following challenge: given national statistics on names in the population, how can we pick a subset of names that best represents the population in terms of traits such as age and gender? McCormick et al. propose to employ a heatmap to adjust the subgroup distribution to the population by trial-and-error. However, this solution is not practical when data is divided into many categories; thus, more scalable solutions are required. Precisely, quoting McCormick et al.: "Another area for future methodological work is formalizing the procedure used to select names that satisfy the scaled-down condition. Our trial-and-error approached worked well here because there were only eight alter categories, but in cases with more categories, a more automated procedure would be preferable". In this direction, we propose to formulate directive 2 as a combinatorial optimization problem over names, whose joint distribution over a set of categories of various traits needs to replicate the population distribution on a smaller scale. In addition, we develop a method for solving the combinatorial problem that is based on quadratic programming.

In this work, we provide an automated approach improving previous methodologies devised to reduce barrier effects. The proposed method can be applied to any population for which name statistics distributed over the categories of various traits are available. Moreover, we empirically show that our approach successfully provides a subgroup of names replicating the population distribution for six countries with very different name statistics: Belgium, Hungary, Poland, Sweden, Switzerland, and

Spain. We also give practical advice for preparing the data. The paper will therefore be useful for other social scientists planning to adopt the NSUM method.

2 Problem Definition

In this section, we provide a mathematical formulation for the problem of selecting a subgroup of names whose joint distribution over categories replicates the population distribution on a smaller scale.

Given a set of n names occurring in a population and a set of m combinations of categories consisting of one or more traits over which the names are distributed (e.g., region and age), each name i has a frequency f_i^j for each possible combination of categorical traits, denoted by the index j. Using the population of Spain as an example, $f_{\text{Juan}}^{(\text{Madrid},20)}$ is the frequency of people named "Juan" whose trait combination is $(\text{Madrid}, 20)$, i.e., who live in "Madrid" and are "20" years old.

Adding all name frequencies f_i^j for one trait combination j provides the frequency of this trait combination in the population, denoted by

$$f^j = \sum_i f_i^j. \tag{1}$$

Consequently, the frequencies of all trait combinations $(f^1, f^2, ..., f^m)$ define the marginal distribution of the trait combinations over the population.

Likewise, adding all trait combination frequencies f_i^j for one name i provides the frequency of this name in the population, denoted by

$$f_i = \sum_j f_i^j. \tag{2}$$

Our approach aims to provide a subset of names N whose joint distribution over the trait combinations replicates the population marginal distribution obtained from computing f^j with equation 1. The subset joint distribution over trait combinations, denoted by $\tilde{f}^j$, is obtained by adding the name frequencies in the subset N and renormalizing, i.e.,

$$\tilde{f}^j = \frac{\sum_{i \in N} f_i^j}{\sum_{i \in N} \sum_k f_i^k}, \tag{3}$$

where the denominator is the normalizing factor resulting from summing all the frequencies over all

trait combinations and names in the subset N.

The resulting problem consists in selecting a subset N such that the joint distribution $\tilde{f}^j$ resembles the original marginal population distribution f^j for each trait combination j. This problem can be formalized as a combinatorial optimization problem minimizing the distance between the two distributions:

$$\arg\min_{N} \quad \sum_{j} D(f^j, \tilde{f}^j), \tag{4}$$

where the space of solutions is conformed by all possible subsets of names N with frequencies constrained in the range $f_i \in [f_{min}, f_{max}]$ in order to satisfy the third McCormick et al.'s directive.

3 Solution Approach

In this section, we describe our solution approach to the problem of finding a subset of names whose joint distribution over trait combinations replicates the population marginal distribution. The section is divided into two parts: first, we discuss the distance measures and their implications for our approach; second, we establish the assumptions under which the problem can be solved with a quadratic programming solver and provide a search technique to overcome the limitations derived from these assumptions.

Our approach is publicly available[1] and can be used by other social scientists aiming to adopt the NSUM method.

3.1 Distance Measure

Choosing a subset of names that jointly represent a population distribution over categorical traits might cause a loss of information if the "scaled-down" distribution diverges from the original one. For this reason, employing a measure that captures this loss is crucial for the goal of our approach. From the information theory perspective, the *Kullback-Leibler* divergence,

$$D_{KL}(f^j, \tilde{f}^j) = f^j \log\left(\frac{f^j}{\tilde{f}^j}\right), \tag{5}$$

[1] https://github.com/coalesce-lab/nsum-name-selection

is defined to capture the loss of information when using a suboptimal distribution, such as $\tilde{f}^j$, instead of the original distribution f_j over all the trait combinations j. Although the Kullback-Leibler quantifies this loss successfully, it suffers from having non-finite values when $\tilde{f}^j = 0$, i.e. when some trait combination is not represented by any name in the subset. For this reason, the *Jensen-Shannon* divergence is preferred,

$$D_{JS}(f^j, \tilde{f}^j) = \frac{1}{2} D_{KL}(f^j, \frac{f^j + \tilde{f}^j}{2}) + \frac{1}{2} D_{KL}(\tilde{f}^j, \frac{f^j + \tilde{f}^j}{2}), \tag{6}$$

which computes the mean of the Kullback-Leibler divergences of f^j and $\tilde{f}^j$ to the average distribution.

Combining the Jensen-Shannon divergence with equation 4 results in a problem hard to solve, in the sense that there do not exist reliable solvers for it. To address this limitation, we consider the *absolute distance*,

$$D_{L1}(f^j, \tilde{f}^j) = \left| f^j - \tilde{f}^j \right|, \tag{7}$$

which allows the formulation of a linear programming problem. Nevertheless, this distance measure is too permissive with high deviations. This is potentially problematic since a high deviation in one category might be preferred instead of several low deviations in different categories. In practice, this might result in one subgroup being under- or over-represented in the subset, thus introducing a strong bias in the surveys. A better option is then the *quadratic distance*,

$$D_{L2}(f^j, \tilde{f}^j) = \left(f^j - \tilde{f}^j \right)^2, \tag{8}$$

which, since it is the square of the absolute distance, it penalizes higher deviations more.

In this work, we use the quadratic distance as our optimization objective. Still, we will evaluate the solutions provided by our approach with respect to the previous distance measures (excluding the Kullback-Leibler divergence because of its non-finite values) to assess the impact of choosing one instead of a theoretically better one.

3.2 Quadratic Programming Reduction

Despite using the quadratic distance, the problem cannot be directly solved as a quadratic programming problem because of the normalizing term in equation 3. In order to overcome this issue, we

define

$$\alpha = \frac{1}{\sum_{i \in N} \sum_k f_i^k}, \tag{9}$$

and we rewrite the objective function so as to obtain a quadratic programming problem:

$$\arg\min_{N} \quad \sum_j \left(f^j - \alpha \sum_i f_i^j x_i \right)^2, \tag{10}$$

where x_i is a binary variable indicating whether $i \in N$.

Equation 10 is a quadratic programming problem under the assumption that α does not depend on N; thus, we need to keep it fixed in order to solve the problem. Although an inaccurate choice of α might lead to suboptimal solutions, some values of α exist for which the solution to the quadratic programming problem is the optimal solution. Moreover, since we select names within the frequency range $[f_{min}, f_{max}]$ and the size of the subset N is fixed as an external problem parameter, $s = size(N)$, with equation 9 we can bound the value of α within the range $[\alpha_{min}, \alpha_{max}]$, where

$$\alpha_{min} = \frac{1}{s \cdot f_{max}}, \tag{11a}$$

$$\alpha_{max} = \frac{1}{s \cdot f_{min}}. \tag{11b}$$

Our approach benefits from the fact that the optimal solution exists within these bounds by searching over λ possible values of α. Then, it solves the corresponding quadratic programming problem ($QP(\alpha)$) for each value of α, providing the best solution.

Algorithm 1 Automated name selection

Input: frequencies $f_i^j \quad \forall i, j$, bounds $[\alpha_{min}, \alpha_{max}]$, and λ
Output: Optimal subset of names N^*

1: $\alpha \leftarrow \alpha_{min}$
2: $D^* \leftarrow inf$ (Set current optimal distance to infinity)
3: **for** $l = 1, \ldots, \lambda$ **do**
4: $N \leftarrow QP(\alpha)$ (Solution to equation 10)
5: $D \leftarrow \sum_j (f^j - \frac{\sum_{i \in N} f_i^j}{\sum_{i \in N} \sum_k f_i^k})^2$
6: **if** $D < D^*$ **then**
7: $N^* \leftarrow N$
8: $D^* \leftarrow D$
9: **end if**
10: $\alpha \leftarrow \alpha + l(\alpha_{max} - \alpha_{min})$
11: **end for**
12: **return** N^*

The whole procedure is presented in algorithm 1, which receives the name frequencies for all trait combinations, the precomputed bounds of α, and the number of search iterations λ. Then, it sets α to the lower bound α_{min}, and the current optimal distance D^* to the maximum possible value (lines 1 and 2 in algorithm 1). Then, the algorithm's main loop starts at line 3 until the end of it. Lines 4 and 5 compute the solution to the quadratic programming problem for a fixed α and the solution distance. Then, line 6 checks if the solution distance improved, and the current optimal solution and the current optimal distance are updated in lines 7 and 8. At the end of the loop, the α is updated by an increment of $l(\alpha_{max} - \alpha_{min})$ to continue the search with the successive α value. Finally, the algorithm terminates after all λ values have been searched, and the optimal solution is returned.

3.3 Limitations

The proposed approach is designed to select a subset of names whose joint distribution replicates, on a smaller scale, the marginal population distribution of trait combinations present in the name statistics. As our model cannot consider traits that do not appear in the data, these traits might not be represented well by the selected names. For example, apart from Spain, none of the data we used in this work gave statistics about the prevalence of the names among immigrants and native citizens. Nevertheless, many names of people with foreign origins tend to differ from local names, and as the migrant population has many origins, most migrant names may have a low overall prevalence, too low to be selected. Combined with barrier effects between migrants and native citizens, this may lead to

an underestimation of network size for immigrants. In practice, a resident with knowledge about the typically used names could identify which names are more common among immigrants and which are more common among native citizens. Adding this information to the data, as a preprocessing step to our proposed approach, could improve the quality of the selected names.

Another possible limitation is that the data may contain variations of the same name, compound names, and names frequently associated with nicknames. This does not directly affect our approach, but it might become another source of error in the surveys when the data are not properly preprocessed. Variations of names that are phonetically similar (e.g., Sarah and Sara) can be summed before using the method (especially when the survey questions are asked by an interviewer, e.g., "How many people do you know whose name is Sara(h)?"). Compound names (i.e., two names that are always used together rather than as a first and middle name, like Jean-Claude or Anna Mae) can be used, unless some people would use it as a compound name and others as a first and middle name, in which case they should be either avoided or grouped (asking about "Mae or Anna Mae"). Names typically associated with nicknames (e.g., Bill for William) could lead to under-recall, as people might not think of the "Bill" they know, when asked to report the number of Williams they know. They should therefore be avoided by eliminating them from the dataset (or alternatively, the occurrence data of William and Bill can be summed, and respondents can be asked explicitly about the two names). If these solutions are still problematic, Equation 10 can be modified to constrain the possible subset N over which the search is performed to subsets not containing problematic names.

In any case, researchers do well to check whether the database contains only first names, first and second names, or the names individuals commonly use, and process the data accordingly. To address the previously mentioned limitations, we recommend involving a person with knowledge about local names and name variations to make the necessary modifications before using our approach.

4 Experimental Analysis

In this section, we apply our approach to the selection of names for six different countries: Belgium, Hungary, Poland, Sweden, Switzerland and Spain. In section 4.1, we describe the data for each country and, in section 4.2, we provide the fitting results achieved by our approach for each country.

Country	Gender	Age	Region	Nationality	Trait combinations	Names in $[0.1, 0.2]$
Poland	2	-	-	-	2	56
Hungary	2	7	-	-	14	63
Belgium	2	3	3	-	18	56
Sweden	2	9	21	-	60 (18 + 42)	96
Switzerland	2	89	4	-	712	116
Spain	2	11	52	2	130 (22 + 104 + 4)	57

Table 1: Structure of the statistics for the six studied countries and trait combinations derived from each. In parentheses the sum of different sets of trait combinations when more than one set was considered.

4.1 Case Studies

The data we used for this paper were derived from six countries. The data for Belgium, Poland, Switzerland and Spain are openly accessible on the websites of Statisitcs Belgium (2023, data from 2021, updated annually), Statistics Poland (2023, data from 2022, updated annually), The Federal Statistical Office of Switzerland (2023, data from 2020, updated annually), and the Spanish Instituto Nacional de Estadistica (2023, data from 2021), respectively. Furthermore, for Sweden and Hungary, we have acquired tailor-made data from Statistics Sweden (data from 2021; only for the population of 18 years and older) and the Hungarian Central Office for Administrative and Electronic Public Service (data from 2021). We chose these countries because they had available name statistics and showed considerable variation in the complexity of these data, making them good test cases for our approach (see Table 1). Total population sizes for each country (in the case of Sweden, of the adult population) were derived from the websites of the same statistical institutions for the same year as the name data.

Table 1 describes the traits appearing in the data for each country, the number of trait combinations, and the number of names in the range $[0.1, 0.2]$. All studied countries divide the name statistics into two genders: male and female. The Polish data we used has no other traits, so it only has two trait combinations. Other countries also divide name statistics into different age groups. In the case of Switzerland, the age division is made per birth year, which results in the largest number of trait combinations (712 when further divided by gender and region). Belgium, Sweden, Switzerland and Spain also provide a division by region, but in the case of Sweden, this division is not mixed with the age range division (i.e., there is a gender × age and a gender × region distribution, but not a gender × age × region distribution). Furthermore, in Spain, data on nationality were available, but not crossed with age and birth region (i.e., gender × age × region distribution and a gender × nationality distribution could be derived, but not a gender × age × region × nationality distribution). In these two cases,

	L1		L2		JS	
Country	**Candidates**	**Solution**	**Candidates**	**Solution**	**Candidates**	**Solution**
Poland	7.08	0.00	25.05	0.00	0.06	0.00
Hungary	61.52	10.14	376.08	18.15	25.67	0.32
Belgium	18.50	0.89	36.91	0.06	0.77	0.01
Sweden	39.64	11.61	63.93	4.69	1.47	0.17
Switzerland	18.69	5.48	2.34	0.12	0.74	0.12
Spain	54.47	26.01	129.16	16.42	2.68	0.69

Table 2: Comparison of the distance between the subsets (candidates and solution) with respect to the population distribution for the six countries. Candidates is the subset of candidate names with frequency in the range $[0.1, 0.2]$ while Solution is the best subset of names provided by our approach. We compare the distance measures: absolute distance (L1), quadratic distance (L2) and Jensen-Shannon divergence (JS).

the procedure was performed simultaneously for the two (three for Spain) sets of trait combinations, and the name subset was chosen that optimized both distributions.

4.2 Results

In this section, we assess the performance of our approach using data from the six countries mentioned earlier. For the evaluation, we chose to limit the size of the subset to 20 names. In addition, we performed 10 search iterations ($\lambda = 10$ in algorithm 1) to show that our approach provides good enough results without the need of extra computation time. Nevertheless, higher values can be used if there is enough time available and more precision is required. In practice, we observed that rarely this leads to better results.

Table 2 summarizes the performance of our model with respect to three different metrics. We see that our approach achieves to reduce all of them considerably. One crucial observation, supported by figure 2, is that not all countries can be approximated with the same precision. The origin of this inaccuracy derives from two different phenomena. The first is when the number of trait combinations is scarce (as for Poland), allowing them to be approximated easily. Scarcity only becomes a problem when the number of available trait combinations fails to capture the heterogeneity in the population successfully. In this case, the only possible solution is adding traits which better divide the population into representative subgroups. The other phenomenon is the possibility that some trait combinations are underrepresented if only names between the range $[0.1, 0.2]$ are considered. In consequence, it is impossible to adjust these trait combinations better. In this case, one can contemplate a wider range of names, e.g., $[0.1, 0.3]$, but this might have other implications on the expected answers in a survey, so it is crucial to value if the accuracy improvement justifies changing these bounds.

To understand how our approach finds a solution, we plotted the distance measures of the solution obtained for each value of α in the third column of figure 2. Our approach starts from the left-most value of α (the smallest one) and proceeds to the higher ones, reporting the best solution found (the one minimizing the quadratic distance). One interesting observation is that the minimum of all the distance measures coincides, except for the absolute distance in the case of Sweden. Therefore, the quadratic distance, while allowing an easier to solve formalization, performs similar to the Jensen-Shannon divergence. The absolute distance also seems a promising choice, since it provides similar performance as the other distances. Nevertheless, we cannot guarantee this will happen in general apart from the six studied countries. For this reason, we still prefer a more robust approach employing the quadratic distance instead of a potentially faster approach using the absolute distance.

Figure 2 shows that for most countries, the solutions represent sets of names whose bearers optimally resemble the population's distribution over the trait combinations. The largest gap is visible for Hungary. Closer inspection of one of the solutions with the most optimal solution for Hungary (see figure 1) shows that various trait combinations are slightly over- or underrepresented in the set of names (penultimate column) compared to the overall population (last column). For instance, men with ages 41-50 make up 8.5% in the total population, but in the subpopulation of bearers of the 20 names, they make up 7%, which is a $\sim 18\%$ relative error. Men with ages 51-60 are also underrepresented, whereas men over 60 years are overrepresented. Women over 50 are also underrepresented. We recommend that users transform the data if the bias is bigger than desired in a particular trait combination, or accumulated over various combinations (e.g., if people over 50 are slightly underrepresented in every region, the accumulated bias could be significant once summing all regions). In the case of Hungary, the data have very fine distinctions for the population of 0-10 years old (3 age categories) and much wider age groups later in life. Combining the three age groups of 10 year-olds and younger may make it easier to obtain a solution that fits the age distribution at both ends. Users could also decide to widen the frequency range $[0.1, 0.2]$ or alter the number of names.

5 Discussion

This work presented a solution approach to select an optimal subset of names for the NSUM method. Our approach aims to reduce the barrier effects derived from non-random mixing in a population; thus, it provides a subset of names which jointly represent the population on a smaller scale. The proposed

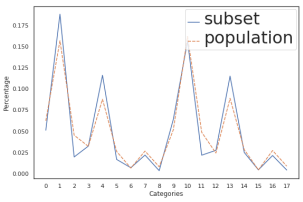
GENDER	AGE	NAMES										Total % in subpopulation with set of 20 names	Total % in population
		FRUZSINA	IZABELLA	JÚLIA	KATA	LÍVIA	MARIANN	RÓZSA	TERÉZ	VALÉRIA	ZITA		
FEMALE	0 yrs	0,066	0,185	0,098	0,027	0,034	0,001	0,003	0,003	0,004	0,049	0,471	0,467
	1-5 yrs	0,387	0,973	0,467	0,213	0,184	0,013	0,020	0,012	0,025	0,273	2,566	2,359
	6-10 yrs	0,453	0,675	0,465	0,374	0,200	0,021	0,031	0,013	0,025	0,278	2,534	2,309
	11-20 yrs	1,182	0,586	0,834	1,324	0,420	0,100	0,074	0,033	0,110	0,579	5,243	4,855
	21-30 yrs	1,576	0,456	0,705	0,915	0,425	0,413	0,101	0,079	0,149	0,641	5,460	5,556
	31-40 yrs	0,556	0,506	0,798	0,700	0,599	1,439	0,191	0,217	0,393	1,037	6,434	6,533
	41-50 yrs	0,117	0,713	0,586	0,206	1,074	1,888	0,375	0,487	0,897	1,819	8,163	8,226
	51-60 yrs	0,010	0,242	0,441	0,025	0,453	0,354	0,648	0,799	1,660	0,812	5,443	6,543
	>60 yrs	0,011	0,499	0,850	0,017	0,515	0,082	2,868	4,116	4,005	0,519	13,482	14,886
MALE	0 yrs	0,016	0,126	0,004	0,011	0,034	0,013	0,171	0,101	0,100	0,023	0,598	0,497
	1-5 yrs	0,086	0,570	0,041	0,062	0,202	0,061	0,701	0,463	0,460	0,158	2,806	2,498
	6-10 yrs	0,090	0,557	0,050	0,064	0,239	0,073	0,510	0,539	0,409	0,200	2,731	2,445
	11-20 yrs	0,181	1,474	0,121	0,190	0,665	0,165	0,835	1,059	0,641	0,550	5,881	5,119
	21-30 yrs	0,272	1,579	0,215	0,338	1,221	0,303	0,515	0,931	0,514	0,689	6,575	5,849
	31-40 yrs	0,514	0,325	0,380	0,615	2,625	0,483	0,562	0,620	0,597	0,871	7,591	6,877
	41-50 yrs	0,812	0,215	0,722	1,224	0,000	0,845	0,732	0,394	0,699	1,384	7,027	8,549
	51-60 yrs	0,758	0,065	0,882	1,452	0,000	0,924	0,147	0,325	0,452	0,355	5,362	6,417
	>60 yrs	1,673	0,031	2,004	3,476	0,000	2,438	0,147	0,969	0,768	0,127	11,633	10,015
		ENDRE	ERIK	ERNŐ	GÉZA	GYÖRGYI	KÁLMÁN	KORNÉL	MÁTYÁS	NÁNDOR	SZILÁRD	100	100

Figure 1: Heatmap for the optimal solution of the Hungarian data. Note: Cells for all columns except the last represent the percentage of people with the given name, gender, and age group in the subpopulation of residents in Hungary who have one of the 20 names as their first name. Cells in the last column represent the percentage of people with the given gender and age group in the total resident population of Hungary.

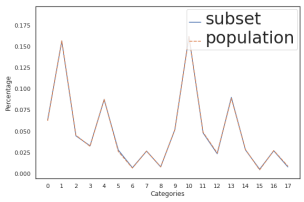
approach receives the population's name statistics and, by formulating a combinatorial optimization problem over names, provides the subset of names that best replicate the population distribution over the categories of different traits, e.g., gender or age. We tested our approach on six countries, showing that it achieves to replicate their diverse population distributions.

Further, we discussed our model's potential limitations, mainly derived from name statistics lacking representative divisions or names presenting diverse variations, compound forms or nicknames. Often, these limitations can be addressed by a citizen familiar with the specific country names. For instance, one could identify names related to migrant populations, as they may be associated with barrier effects. Concerning name variations, these can also be addressed by a citizen checking each name's variations. Otherwise, removing problematic names from being considered by our approach is always possible.

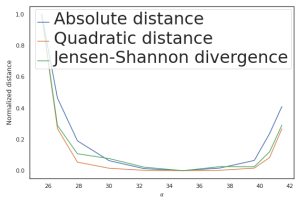
The results show that the proposed approach achieves to replicate the population distribution on a smaller scale for the six studied countries. Thus, our approach can robustly be applied to countries with diverse name statistics. Moreover, the approach can be extended to different domains involving the selection of subsets whose joint distribution replicates a target distribution. For instance, social capital researchers often use so-called "position generators" to ask survey respondents whether they know people in occupations with different levels of prestige and income, to estimate their access to social capital. If data on the income and gender distribution of each occupation in a society are



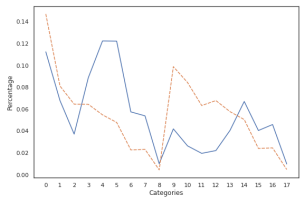
(a) Belgium candidates



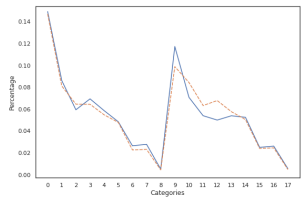
(b) Belgium solution



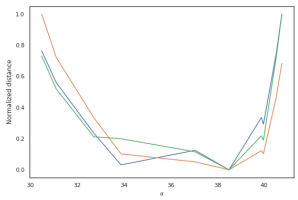
(c) Belgium solution distance



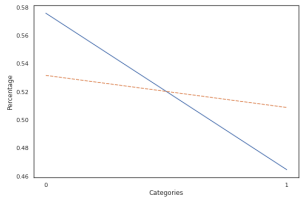
(d) Hungary candidates



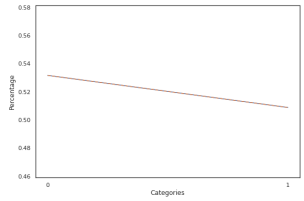
(e) Hungary solution



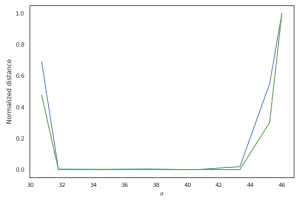
(f) Hungary solution distance



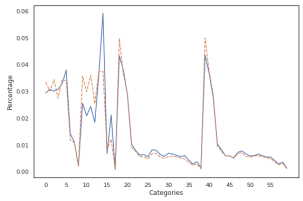
(g) Poland candidates



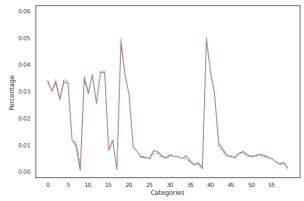
(h) Poland solution



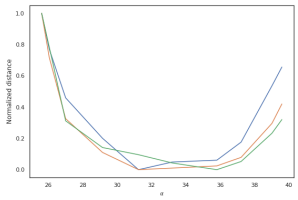
(i) Poland solution distance



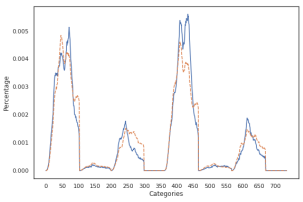
(j) Sweden candidates



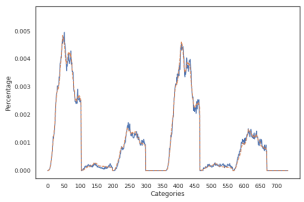
(k) Sweden solution



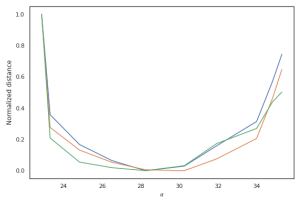
(l) Sweden solution distance



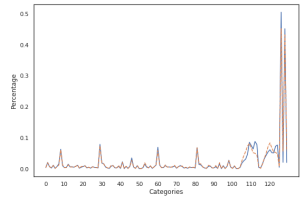
(m) Switzerland candidates



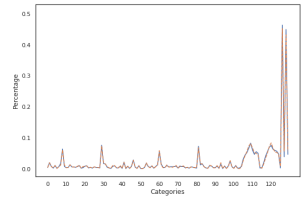
(n) Switzerland solution



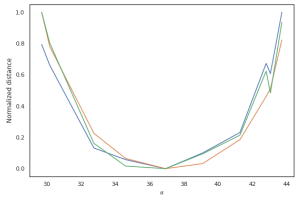
(o) Switzerland solution distance



(p) Spain candidates



(q) Spain solution



(r) Spain solution distance

Figure 2: Candidates subset distribution (names in range $[0.1, 0.2]$) compared to the original population distribution (first column), solution subset distribution compared to the original population distribution (second column), and solution distance for each α for the absolute distance, the quadratic distance and the Jensen-Shannon divergence (third column). Each row corresponds to a different country: Belgium, Hungary, Poland, Sweden, Switzerland and Spain.

available, researchers could adopt the method to select a set of occupations that together represent the overall population's occupational structure over these traits. Doing so may avoid inadvertently selecting occupations biased toward one gender, and attributing resulting gender differences to a lack of social capital rather than homophily in acquaintanceship patterns.

Acknowledgments

The authors are grateful for the funding of the La Caixa Foundation (SR0587) for the research project "BRIDGES. A Network Science Approach to Social Cohesion". Miranda Lubbers is further grateful for the support received from the ICREA Acadèmia program.

References

Bernard, H. R., Hallett, T., Iovita, A., Johnsen, E. C., Lyerla, R., McCarty, C., Mahy, M., Salganik, M. J., Saliuk, T., Scutelniciuc, O., et al. (2010). Counting hard-to-count populations: The network scale-up method for public health. *Sexually Transmitted Infections*, *86*(Suppl 2), ii11–ii15.

Bernard, H. R., Johnsen, E. C., Killworth, P. D., & Robinson, S. (1991). Estimating the size of an average personal network and of an event subpopulation: Some empirical results. *Social Science Research*, *20*(2), 109–121.

DiPrete, T. A., Gelman, A., McCormick, T., Teitler, J., & Zheng, T. (2011). Segregation in social networks based on acquaintanceship and trust. *American Journal of Sociology*, *116*(4), 1234–83.

Feehan, D. M., & Salganik, M. J. (2016). Generalizing the network scale-up method: A new estimator for the size of hidden populations. *Sociological Methodology*, *46*(1), 153–186.

Instituto Nacional de Estadistica. (2023, January). *Apellidos y nombres más frecuentes*. https://www.ine.es/dyngs/INEbase/es/operacion.htm?c=Estadistica_C&cid=1254736177009&menu=resultados&idp=1254734710990

Lubbers, M. J., Molina, J. L., & Valenzuela-Garcia, H. (2019). When networks speak volumes: Variation in the size of broader acquaintanceship networks. *Social Networks*, *56*, 55–69.

Maltiel, R., Raftery, A. E., McCormick, T. H., & Baraff, A. J. (2015). Estimating population size using the network scale up method. *The Annals of Applied Statistics*, *9*(3), 1247.

McCarty, C., Killworth, P. D., Bernard, H. R., Johnsen, E. C., & Shelley, G. A. (2001). Comparing two methods for estimating network size. *Human Organization*, *60*(1), 28–39.

McCormick, T. H., Salganik, M. J., & Zheng, T. (2010). How many people do you know?: Efficiently estimating personal network size. *Journal of the American Statistical Association*, *105*(489), 59–70.

Rivera, M. T., Soderstrom, S. B., & Uzzi, B. (2010). Dynamics of dyads in social networks: Assortative, relational, and proximity mechanisms. *Annual Review of Sociology*, *36*, 91–115.

Statisitcs Belgium. (2023, January). *Male and female first names*. https://statbel.fgov.be/en/themes/population/family-names-and-first-names/male-and-female-first-names

Statistics Poland. (2023, January). *Lista imion występujących w rejestrze PESEL*. https://dane.gov.pl/en/dataset/1667,lista-imion-wystepujacych-w-rejestrze-pesel-osoby-zyjace

The Federal Statistical Office of Switzerland. (2023, January). *First and last names in Switzerland*. https://www.bfs.admin.ch/bfs/en/home/statistics/population/births-deaths/names-switzerland.html