Estimating the Total Volume of Queries to Google

Fabrizio Lillo
University of Bologna
Bologna, Italy
fabrizio.lillo@unibo.it

Salvatore Ruggieri
University of Pisa
Pisa, Italy
salvatore.ruggieri@unipi.it

ABSTRACT

We study the problem of estimating the total volume of queries of a specific domain, which were submitted to the Google search engine in a given time period. Our statistical model assumes a Zipf's law distribution of the population in the reference domain, and a non-uniform or noisy sampling of queries. Parameters of the distribution are estimated using nonlinear least square regression. Estimations with errors are then derived for the total number of queries and for the total number of searches (volume). We apply the method on the *recipes and cooking* domain, where a sample of queries is collected by crawling popular Italian websites specialized on this domain. The relative volumes of queries in the sample are computed using Google Trends, and transformed to absolute frequencies after estimating a scaling factor. Our model estimates that the volume of Italian recipes and cooking queries submitted to Google in 2017 and with at least 10 monthly searches consists of 7.2B searches.

CCS CONCEPTS

• **Information systems → Web search engines**; • **Mathematics of computing → Probabilistic inference problems**.

KEYWORDS

Search engine query; Volume estimation; Zipf's law; Google Trends

ACM Reference Format:
Fabrizio Lillo and Salvatore Ruggieri. 2019. Estimating the Total Volume of Queries to Google. In *Proceedings of the 2019 World Wide Web Conference (WWW '19), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3308558.3313535

1 INTRODUCTION

The problem of computing the total number of searches (volume) of queries belonging to a specific domain is extremely relevant and, at the same time, challenging. From a business perspective, the total volume $\mathcal{V}$ of queries quantifies the potential market of search engine advertising in the domain. An even more interesting quantity is the total volume $\mathcal{V}_v$ of queries searched at least v times. $\mathcal{V}_v$ quantifies the potential market of queries worth to bid on. Related to the above, the total number of queries N in the domain, or of queries N_v searched at least v times, are also gold nuggets. However, the stream of queries submitted to a search engine is so massive that it is impractical to keep frequency counts of every possible query, particularly of those in the long tail of the distribution.

WWW '19, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6674-8/19/05.
https://doi.org/10.1145/3308558.3313535

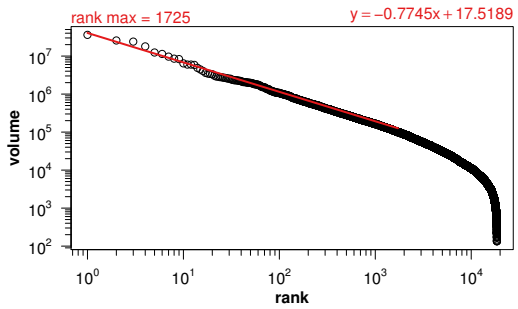

Figure 1: Empirical rank-volume distribution (scaled Google Trends estimates). Best view in color.

Here we study the problem of estimating the total volume of queries submitted to the Google search engine for a specific domain in a given time period. While our method is in principle general, in this paper we apply it to data in the domain of *recipes and cooking*. Such a domain consists of queries with the name of the recipe of a dish, excluding drinks. The advantage over other domains is that it is relatively easy to collect sample recipes and to validate whether a given text is a recipe or not. In particular, we crawled popular websites of Italian recipes and cooking, collecting a sample of more than 120K queries. We then resorted to Search Engine Optimization (SEO) tools, and in particular to Google Trends[1], for obtaining estimates of the volume of each query in the sample for the whole year 2017.

The motivation for the model adopted in this paper comes from the evidence of Figure 1, which shows the empirical rank-volume distribution obtained using estimates of Google Trends. Actually, Google Trends provides relative volumes, not absolute frequencies, thus to find absolute volumes we need to estimate an appropriate scaling factor. This is done by correlating relative volume with ground truth continuous data. We rely on query impression summaries provided by the Google Search Console of a top-ranked website. Indeed Figure 1 reports absolute volumes obtained by rescaling relative ones. The most difficult task in our problem is to estimate the volume of the queries in the population which do *not* belong to the empirical sample. For this reason we do a precise statistical assumption on the rank-volume distribution of the whole population of queries (i.e. observed and unobserved). Our statistical model assumes a Zipf's law distribution of the population, as suggested by the empirical distribution of Figure 1 and previous related work [16]. In order to cope with computational issues, SEO tools may adopt sampling strategies and/or approximated counting techniques, e.g., *count-min sketch* summaries [6, 8], that favor volume

[1] https://trends.google.com

estimation of popular queries against the ones in the long tail of the distribution. This yields the visible drop in volume in the tail of the empirical distribution of Figure 1, with only 18.5K queries being assigned a non-zero volume estimate by Google Trends. We are able to model this behavior by assuming that empirical sampling from the population is not uniform, but it depends on the true rank of a query (*non-uniform* sampling). Moreover, in order to account for approximations in the SEO tool data, we additionally assume that the estimates are noisy, and discuss two specific sampling schemes (*noisy* and *sketchy* sampling). Parameters of the Zipf distribution are estimated using Nonlinear Least Square (NLS) regression. Simulations show such estimators perform better than an alternative approach based on Power law parameter estimation. We derive then estimators of total volumes $\mathcal{V}$ and $\mathcal{V}_v$, and total number of queries N and N_v, including closed formula for statistical errors of such estimators.

In summary, this paper makes the following contributions:

- we formalize the problem of estimating the total volume of queries submitted to a search engine, propose a statistical model which is consistent with empirical data, and infer parameters of the statistical model that perform well under simulated conditions;
- we design a procedure for estimating relative volumes of a set of queries that overcomes the rounding error introduced by Google Trends, and devise a statistical model for scaling relative volumes to absolute ones starting from ground truth SEO data;
- we apply the approach to the domain of recipes and cooking for queries in Italian, and produce estimations for the volume $\mathcal{V}_v$ of queries searched at least v times in 2017.

This paper is organized as follows. First, we report on related work in Section 2. Next, Section 3 states the main problem by modelling the rank-volume distribution of queries as Zipf's law. Section 4 first discusses the impact of non-uniform sampling from a Zipf's law, which is consistent with empirical data. Then, estimators of the parameters of the Zipf's law are introduced, and adopted for estimating the number and total volume of queries in the population. Section 5 describes the approximation introduced by computing relative volumes from Google Trends data, and presents a statistical model for scaling relative volumes to absolute ones. Section 6 describes the available empirical data obtained by collecting Google Trends relative volumes, and applies the scaling method of Section 5 and the estimators of Section 4 to the empirical data. Conclusions summarize the contribution of the paper.

2 RELATED WORK

Pareto distributions and Zipf's laws are ubiquitous in empirical data of many fields [5, 14], and in information retrieval in particular [16]. Several works [2, 3, 10, 16] have observed that the probability that a query is searched v times in a query log is approximately Power law distributed, namely $P(V = v) \propto 1/v^{\alpha}$. This implies (see e.g., [1, 4]) that the the probability that a query is ranked i-th follows a Zipf's law $P(R = i) \propto 1/i^{\beta}$ for $\beta = 1/(\alpha - 1)$. This information on query frequencies/ranks has been used to optimize caching and distribution strategies in search engines and peer-to-peer systems.

There is a huge literature on the estimation of parameters of Power law distributions and Zipf's law. Popular methods [16] have relied on: graphical methods, straight-line approximation, maximum-likelihood estimation. The estimated tail exponent, even in simulated data, significantly depends on the adopted method [12]. A major breakthrough was the method proposed in [5], which consists in a maximum-likelihood estimation, with a cutoff for the fitting region determined with a Kolmogorov-Smirnov test. This method is implemented in the `powerLaw` package [11] of R, which we used extensively in our analyses.

A related stream of literature considers the *unseen species* problem. As originally stated, the problem asks how many biological species are present in a region, given that in an observation campaign a certain number of species with their relative frequency have been observed. In our case, the problem is that we have (noisy) estimates of the frequency of a certain number of queries, and we want to estimate the number of unobserved queries and their frequency. Despite there are several estimators for the unseen species problem (for example, the Good-Toulmin estimator and its extensions [15]), the problem tackled here is different in an important aspect. In the unseen species problem, it is often assumed that in the sample used to build the estimator, the observed frequencies are proportional to the true frequencies in the population. In other words, there is no bias in the construction of the sample. In our approach, the elements of the sample are chosen *ex-ante* and not necessarily the probability of being in the sample is proportional to true frequency.

Google Trends has been widely used for correlating search trends with offline indicators of economic activity, business performance, disease spreading, brand value and awareness, box-office revenue and audience, stock market variability, etc. [17] presents a brief review of the literature. To the best of our knowledge, all works make use of relative volumes only. Their conclusions are stated in relative terms, such as increase/decrease of a searched topic. Here, we first attempt at determining absolute volumes of sample queries, and at inferring how they aggregate over all queries in a domain.

In general, there is little documentation on how SEO tools collect query logs for providing estimates of search frequencies. Google Trends and Google AdWords can rely on Google search engine logs. Similarly for services provided by other search engines. Independent SEO tools (Searchvolume.io, Ubersuggest, Semrush, Keywordkeg, etc.) rely on a more limited user base. [17] compares Google Trends and Baidu Index (restricted to searches from China only), and finds that their estimates are highly correlated. An advantage of Baidu Index over Google Trends is that it provides absolute estimates, not relative ones. For reference domains restricted to searches from China, by using Baidu Index instead of Google Trends, one would save the task of scaling relative to absolute volumes described in Section 5.

3 PROBLEM STATEMENT

Let us assume the population of queries in the reference domain is composed by N queries, and that the rank-volume distribution of such population follows a Zipf's law. Formally, the volume V_i of the i-th most popular query q_i, for $i \in [1, N]$, is:

$$V_i = \frac{c}{i^{\beta}}. \tag{1}$$

The parameters c and β are called the *intercept* and the *coefficient* respectively. The total volume over the population is thus:

$$\mathcal{V} \equiv \sum_{i=1}^{N} \frac{c}{i^{\beta}} = c[\zeta(\beta) - \zeta(\beta, N+1)] \tag{2}$$

where $\zeta(x)$ and $\zeta(x, y)$ are the Riemann zeta and Hurwitz functions, respectively. If N, c, and β are known one can easily determine $\mathcal{V}$. As discussed in the introduction, however, there are several reasons that make this impossible in practice. The problem that we investigate in this paper consists of estimating $\mathcal{V}$ starting from an empirical sample of volumes $v_1, \ldots, v_n$, for $n < N$ sample queries. Without any loss of generality, we assume that the observations are ranked, namely $v_1 \geq v_2 \geq \ldots \geq v_n$.

The problem can be decomposed in two parts: (1) since true absolute volumes V_i are not observable, even for the subset of n queries, we propose a method for estimating them; (2) having a possibly noisy estimation v_i for V_i in a possibly non-uniform sample subset, we consider the problem of estimating the total query volume $\mathcal{V}$, including also the volume of the unobserved queries. Problem (2) is tackled first in the next section, while problem (1) is discussed in Section 5.

4 MODELLING AND ESTIMATION

4.1 Sampling from a Zipf

Starting from the assumption that the volume of the query population follows a Zipf distribution (see Eq. 1), we observe that the empirical distribution in Figure 1 shows a drop of volume in its tail. We intend here to investigate on this. We will consider the effects of different sampling methods from a Zipf's law, and check whether the conclusions are consistent with our empirical data.

Clearly, uniform sampling from a Zipf's law cannot explain the drop of volume in the tail of the empirical distribution. In fact, queries in an empirical sample are rarely chosen uniformly. The approach followed in our reference domain, for instance, relies on collecting recipe names from specialized websites. These typically conduct a keyword research effort in targeting high-volume keywords. As a consequence, our empirical data suffers from an unavoidable selection bias in favor of high-volume queries. A similar bias against very low volume queries is introduced by SEO tools (e.g., Google Trends) used to obtain volume estimates of queries in a sample. In summary, our empirical data is likely to be a non-uniform sampling of the query population. We assume here that sampling depends on the true rank, and call this *non-uniform sampling*. Formally, we assume that the i-th query q_i is sampled with a probability $p(i)$. We want to check whether the observed rank plot obtained by a sample of the population is different from a Zipf's law. To this end, we consider a geometric sampling $p(i) \propto p(1-p)^{i-1}$, i.e., the sampling probability decays exponentially with the rank. For example, if $p = 0.01$, the probability that the query with the largest volume in the population is observed is p, then the second, third, fourth, etc. query in terms of volume will be observed (i.e. sampled) with probability $0.99p$, 0.99^2p, 0.99^3p, etc. Figure 2 shows a numerical simulation with the following parameters[2] of

[2]The choice of β, in particular, has been driven by the empirical distribution of Figure 1.

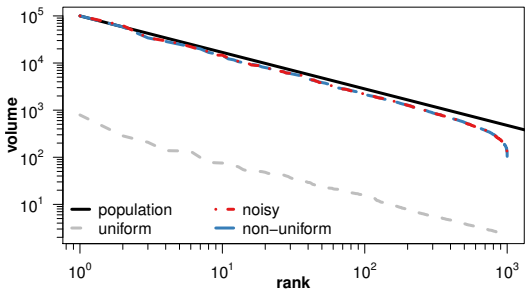

Figure 2: Simulation of sampling from a Zipf's law.

the population:

$$N = 10^6, \quad c = 10^5, \quad \beta = 0.7745. \tag{3}$$

Samples consist of $n = 1000$ queries, and $p = 0.001$ is set for the geometric sampling. The black line is the whole population, the blue line is obtained with geometric sampling while the grey line is obtained with uniform sampling. The non-uniform sampling is consistent with the tail of the empirical distribution in Figure 1.

As a second aspect worth to be considered, we have mentioned that SEO tools typically provide approximated values of the true volume of queries, due to their sampling strategy and computational heuristics in frequency counting. Another source of approximation will be discussed in Section 5.1. Therefore, our empirical data is drawn from noisy values X_i's of the true V_i's. We assume that:

$$X_i = V_i \epsilon_i$$

where ϵ_i are independent noise with common distribution characterized by the same mean μ and variance σ_i^2. Clearly, the presence of noise scrambles the frequencies, thus the most frequent according to X is not necessarily the most frequent according to V. Figure 2 includes also a noisy and non-uniform sample (red line) generated assuming ϵ_i normally distributed, but truncated to 0 to avoid negative V_i's. Parameters are set as follows: $\mu = 1$, i.e., noise is unbiased, and $\sigma_i^2 = 0.01/9$, i.e., 99.7% of noise is in the range $\pm 3\sigma = \pm 10\%$ of the true value. Noisy and non-uniform sampling (hereafter *noisy sampling*) produces an empirical distribution very close to the one of non-uniform sampling and that is also consistent with our empirical data.

A yet another way to model computationally approximated counting, as provided by count-min sketches [6, 8], is to set:

$$X_i = V_i + \gamma_i c$$

where γ_i is uniformly distributed in the range $[0, \gamma]$. In such case, the noise overestimates V_i up to a fraction γ of the top volume $V_1 = c/1^{\beta} = c$. For low volumes, the noise may considerably increase the observed value. However, for a sufficiently low γ, the non-uniform sampling alleviates from this problem, since low volumes are sampled with low probability. We set $\gamma = 0.001$ in simulations. The empirical distribution generated lies in between the ones of non-uniform and noisy sampling. For readability reasons, it is not shown in Figure 2. We call such model the sketchy and non-uniform sampling, hereafter *sketchy sampling*.

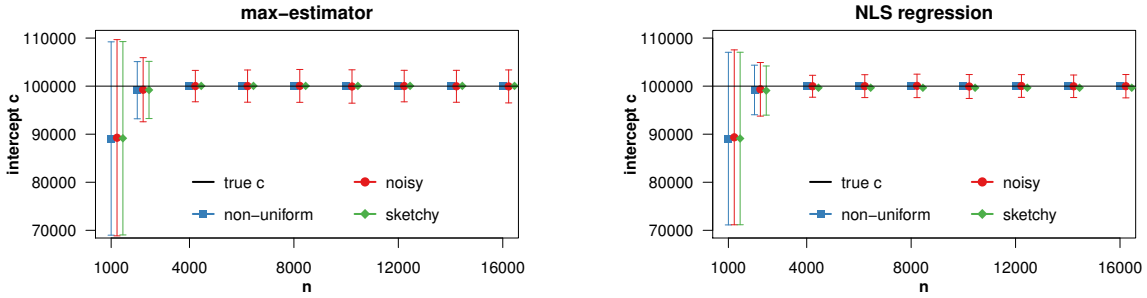

Figure 3: Simulations on estimation of c: error bars (mean ± stdev).

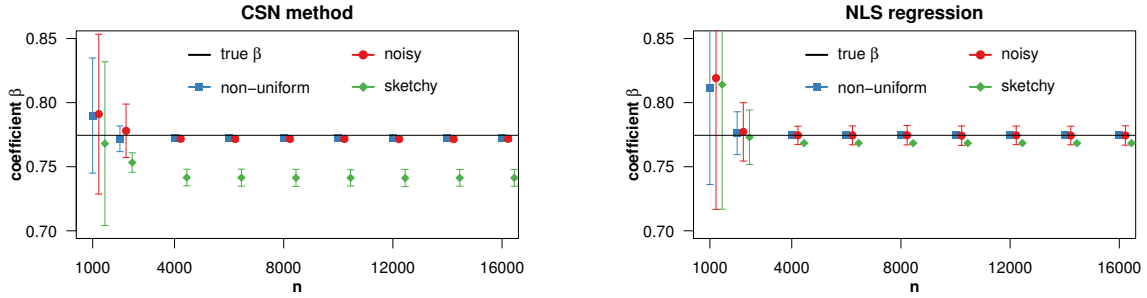

Figure 4: Simulations on estimation of β: error bars (mean ± stdev).

4.2 Estimating β and c

We consider now the estimation of the coefficient β and intercept c in Eq. 1 by exploring two alternative methods for each of them. Regarding the coefficient, we observe that β is the same coefficient of the p.d.f. of the continuous Zipf's law:

$$p_i = \frac{1}{\zeta(\beta)\ i^{\beta}}$$

Thus, we can use the well-known method of Clauset, Shalizi and Newman [5] (hereafter, the CSN method) for estimating the β parameter in Eq. 1. Strictly speaking, [5] is a maximum-likelihood estimator $\hat{\alpha}$ of the α exponent of the Power law of volume distribution, $P(V = v) \propto 1/v^{\alpha}$, from the high-volume tail of empirical data $v_{max} \leq \ldots \leq v_1$. Since in many empirical data the Power law tail is observed only for a range of values, [5] uses a Kolmogorv-Smirnov like test to determine v_{max} which is the optimal value after which the distribution is Power law tailed. Using the well-known relation $\beta = 1/(\alpha - 1)$ between exponents of Power law and continuous Zipf's law (see [1, 4]), we obtain the estimate $\hat{\beta} = 1/(\hat{\alpha} - 1)$ of the coefficient of the rank-volume distribution for top ranks 1 to *max*. The theoretical advantage of this method is that it automatically selects the rank *max* from which to regress the coefficient.

The second estimator of β is to use standard Nonlinear Least Square (NLS) regression of the volume V_i from the rank i. This means that the parameters c and β are those minimizing the sum of squares: $\sum_{i=1}^{M} \left(V_i - \frac{c}{i^{\beta}}\right)^2$, where M is the maximal rank considered in the regression[3]. Since the empirical data follows such distribution for the top ranks, we regress only the top $M = max$ rank-volume data, where *max* is the rank returned by the CSN method. NLS has two advantages over CSN. First, intercept c and coefficient β are estimated together in the same procedure. Second, the regression directly estimates β, while in the CSN method β is estimated with a formula involving the estimator of α. Finally, the second estimator of the intercept that we consider here is the maximum observed value, namely v_1. We call it the *max-estimator* of c. This is motivated by observing that $V_1 = c/1^{\beta} = c$, namely the intercept c is the volume of the top ranked query in the population.

Let us now investigate how these estimators are affected by the non-uniform, noisy, and sketchy sampling from a Zipf's law. Numerical simulations with parameters as in (3), are repeated at the variation of the sample size n for 1000 times and results averaged.

Figure 3 shows that both the max-estimator and the NLS regression converge to the true value of the intercept c. For noisy data, however, there is some error, which is proportional to the noise level (set to ±10%). Variability is slightly lower for NLS regression. Larger error bars can be observed for small values of n. They are due to the chances of not having the highest volume of the population included in the sample. This chance is controlled by the $p = 0.001$ parameter in the geometric sampling. Smaller values lead to larger standard deviation, and, symmetrically, larger values to smaller standard deviation. Thus, in practical settings, the selection of the sample queries must carefully consider the issue of including the most popular queries in the empirical sample. This has been

[3]NLS regression requires to specify initial values for β and c to start with. We compute them using ordinary (linear) least squares (OLS) of the log, i.e. minimizing $\sum(\log V_i - \log c - \beta \log i)^2$. This method cannot be used as an alternative to NLS since it gives too much importance to deviations of low rank queries with respect to high rank queries.

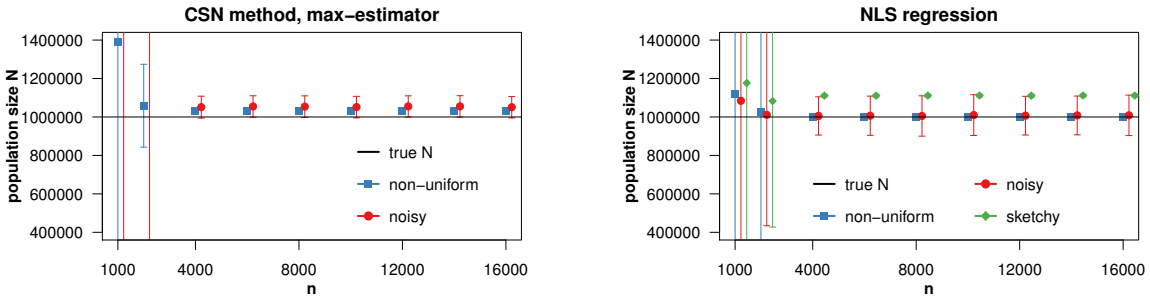

Figure 5: Simulations on estimation of the population size N: error bars (mean ± stdev).

one of our main concerns in collecting queries in the recipe and cooking domain.

Figure 4 shows some differences in the estimation of β. Regarding the CSN method, the estimated values for non-uniform and noisy samplings are slighly lower than the true β. Underestimation in the sketchy sampling case is, instead, considerable. Regarding the NLS regression, it is unbiased for non-uniform and noisy sampling. For sketchy sampling, β is slightly underestimated. Estimations rapidly converge for increasing n's, except for noisy sampling in the case of NLS, and for sketchy sampling in the case of CSN.

Finally, all estimations are weakly dependent on n: starting for samples of 0.4% of the population, they become stable.

4.3 Estimating N

In the following, we will focus on a simple but effective estimator of the size N of the query population. We assume to know V_N, namely the smallest volume of a query in the population. This assumption is realistic for absolute frequencies, since $V_N \simeq 1$. From Eq. 1, for $i = N$, we have $N = (c/V_N)^{1/\beta}$. This motivates the following estimator:

$$\hat{N} = \left(\frac{\hat{c}}{V_N}\right)^{\frac{1}{\hat{\beta}}} \tag{4}$$

where $\hat{c}$ is an estimator of c and $\hat{\beta}$ is an estimator of β. Eq. 4 can be extended to an estimator of the number of queries whose volume is greater or equal than a given value v as:

$$\hat{N}_v = (\hat{c}/v)^{1/\hat{\beta}} \tag{5}$$

Numerical simulations with parameters as in (3) are shown in Figure 5 for: (1) $\hat{\beta}$ obtained by the CSN method and $\hat{c}$ obtained by the max-estimator; and (2) $\hat{\beta}$ and $\hat{c}$ obtained by NLS regression. The first method is biased, showing a slight overestimation for non-uniform and noisy sampling and a large overestimation for sketchy sampling (not shown because exceeding the y-axis limits). The second method converges to the true value of N for non-uniform and noisy sampling (on average), and it slightly overestimates it for sketchy sampling. These findings are intuitive. They follow from Eq. 4, by observing that, if $\hat{c}$ is unbiased (as shown in Figure 3), then the estimator $\hat{N}$ has a bias proportional to the power of $1/\hat{\beta}$. We know from Figure 4 that $\hat{\beta}$ underestimates β for the CSN method or for the sketchy sampling. The only advantage of first method over the second one, is a smaller variability of the estimates in the case of noisy sampling. Again, this is a direct consequence of the smaller variability of β estimates (see Figure 4).

4.4 Estimating $\mathcal{V}$

Building on the estimators and simulations conducted so far, the proposed procedure for estimating the total volume $\mathcal{V}$ is composed of the following steps:

- estimate β and c, as described in Section 4.2;
- use the estimated $\hat{\beta}$ and $\hat{c}$ as inputs for estimating N as shown in Section 4.3;
- the estimator of $\mathcal{V}$ is obtained from Eq. 2 as follows:

$$\hat{\mathcal{V}} = \hat{c}[\zeta(\hat{\beta}) - \zeta(\hat{\beta}, \hat{N} + 1)]$$

Notice that by Eq. 4, the estimator $\hat{\mathcal{V}}$ can be stated using only $\hat{\beta}$ and $\hat{c}$:

$$\hat{\mathcal{V}} = \hat{c}[\zeta(\hat{\beta}) - \zeta(\hat{\beta}, \left(\frac{\hat{c}}{V_N}\right)^{\frac{1}{\hat{\beta}}} + 1)] \tag{6}$$

These estimators can be generalized to estimators of the total volumes of queries with minimum volume v by replacing V_N by v:

$$\hat{\mathcal{V}}_v = \hat{c}[\zeta(\hat{\beta}) - \zeta(\hat{\beta}, (\hat{c}/v)^{1/\hat{\beta}} + 1)] \tag{7}$$

Let us continue the previous numerical simulations. With the settings in (3), it turns out $\mathcal{V} = 9,609,224$. First consider using the NLS regression method in the first step of the procedure. Figure 6 shows that $\hat{\mathcal{V}}$ converges to $\mathcal{V}$ for non-uniform and noisy sampling, and overestimates it for sketchy sampling. For noisy sampling, there is some variability, which is in the order of the noise introduced during sampling (±10%). The overestimation in the case of sketchy sampling follows from the overestimation of N (see Figure 5).

Consider now the case of using in the first step of the procedure the CNS method coupled with the max-estimator. The total volume is slightly overestimated for non-uniform sampling and for noisy sampling. In the latter case, there is some variability, which appears lower than for the CSN method. This can be tracked back to lower variability in the estimation of β (see Figure 4). For sketchy sampling, the overestimation is very large: it is out of the bounds of the plot in Figure 6. Again, this can be traced back to a larger underestimation of β compared to the CSN method.

The impact of biased $\hat{\beta}$ on the estimated total volume $\hat{\mathcal{V}}$ can be readily explained when $\hat{c} = c$ – which holds in simulations, as shown in Figure 3. In Figure 7 we plot Eq. 6 as a function of $\hat{\mathcal{V}}$, under the assumption that V_N is known. The left plot shows

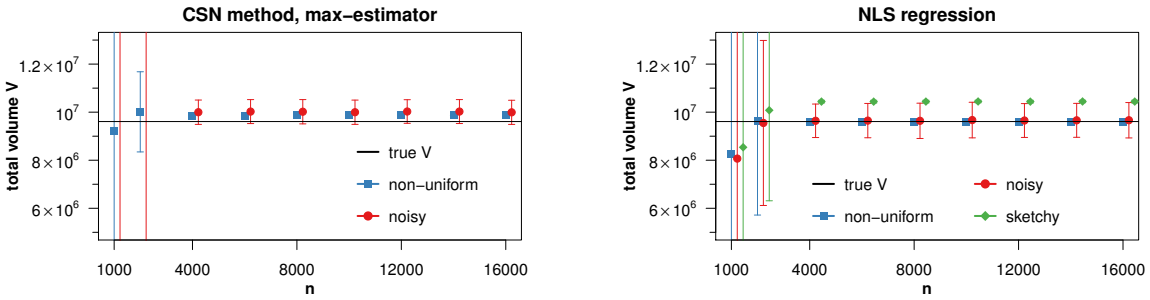

Figure 6: Simulations on the estimation of the total volume $\mathcal{V}$: error bars (mean ± stdev).

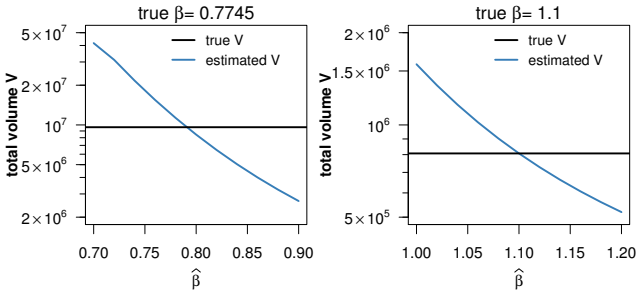

Figure 7: Estimated volume (Eq. 6) as a function of $\hat{\beta}$, assuming $\hat{c} = c$. Simulation parameters: $N = 10^6$, $c = 10^5$, β in title.

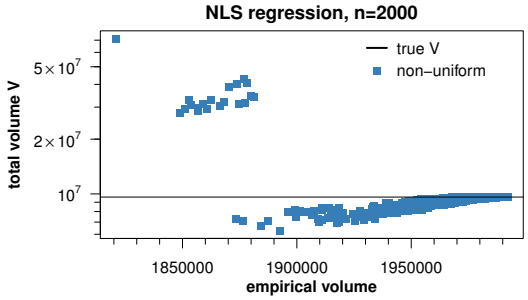

Figure 8: Scatterplot of empirical vs estimated total volume.

simulations for the parameters in (3) used so far. The right plot uses the same N and c, but a β greater than 1. In both cases, the bias of $\hat{\mathcal{V}}$ is inversely proportional to bias of β. Note the log scale in the y-axis, which comes from the fact that β appears as exponent in Eq. 6. For β's lower than 1, error (or variability) of the estimator $\hat{\beta}$ has a greater impact on error (or variability) of $\hat{\mathcal{V}}$ than for β's greater than 1.

In all three sampling models, the performances become stable from $n = 4,000$ on, which is 0.4% of the population. Let us now focus on small sample sizes, for which instead there is a large standard deviation over the experimental runs. Fix $n = 2,000$, and consider NLS regression and uniform sampling. The estimated $\hat{\mathcal{V}}$ is approximately $\mathcal{V} \pm 3.9 \times 10^6$, i.e., the standard deviation is 4 times the (unbiased) average. What is the source of such variability? Figure 8 shows the scatter plot of empirical volume vs estimated total volume over the 1,000 experimental runs. Runs where the generated sample has a low total empirical volume exhibit most of the variability (notice that the y-axis is in logscale). If the total empirical volume is sufficiently large, even small samples converge to the true volume. This reinforces our previous conclusion that, in practical settings, the selection of sample queries must carefully include popular ones, especially for small size samples.

As a summary of the simulations, we therefore recommend using the NLS regression method for estimating c and β, and, using Eqs. 6–7, for estimating $\mathcal{V}$ and $\mathcal{V}_v$.

4.5 Errors on the estimates

We now compute the error on the estimated N obtained from Eq. 4. Using the propagation of errors under the assumption that the errors on β and c are independent, the error on $\hat{N}$ is:

$$\Delta N = \sqrt{\left(\frac{\partial \hat{N}}{\partial \hat{c}}\Delta c\right)^2 + \left(\frac{\partial \hat{N}}{\partial \hat{\beta}}\Delta \beta\right)^2}$$

To have a more conservative estimate of ΔN, taking into account correlations between errors, one can replace the previous formula with the sum of the absolute values:

$$\Delta N = \left|\frac{\partial \hat{N}}{\partial \hat{c}}\Delta c\right| + \left|\frac{\partial \hat{N}}{\partial \hat{\beta}}\Delta \beta\right| \tag{8}$$

The partial derivatives in the previous expressions are:

$$\frac{\partial \hat{N}}{\partial \hat{c}} = \left(\frac{\hat{c}}{V_N}\right)^{\frac{1}{\hat{\beta}}} \frac{1}{\hat{\beta}\hat{c}} = \frac{\hat{N}}{\hat{\beta}\hat{c}}$$

$$\frac{\partial \hat{N}}{\partial \hat{\beta}} = -\left(\frac{\hat{c}}{V_N}\right)^{\frac{1}{\hat{\beta}}} \frac{1}{\hat{\beta}^2} \log\frac{\hat{c}}{V_N} = -\frac{\hat{N}}{\hat{\beta}^2} \log\frac{\hat{c}}{V_N}$$

From these values and the knowledge of Δc and $\Delta \beta$ (obtained from the NLS regression), it is possible to compute ΔN.

The computation of the error on the total volume is a bit more involved. Consider $\hat{\mathcal{V}}$ as a function of $\hat{c}$ and $\hat{\beta}$ (see Eq. 6). To find the error on $\hat{\mathcal{V}}$ we compute its derivatives with respect to $\hat{c}$ and $\hat{\beta}$. We find:

$$\frac{\partial \mathcal{V}}{\partial \hat{c}} = \frac{\hat{\mathcal{V}}}{\hat{c}} + \hat{N}\zeta\left(\hat{\beta}+1, \hat{N}+1\right)$$

$$\frac{\partial \mathcal{V}}{\partial \hat{\beta}} = \hat{c}\left(\zeta'(\hat{\beta}) - \zeta^{(1,0)}\left(\hat{\beta}, \hat{N}+1\right) - \frac{\hat{N}\log\left(\frac{\hat{c}}{V_N}\right)\zeta\left(\hat{\beta}+1, \hat{N}+1\right)}{\hat{\beta}}\right)$$

where $\zeta'(x)$ is the derivative[4] of the Riemann Zeta function and $\zeta^{(1,0)}(s,a)$ is the partial derivative of the Hurwitz function with respect to s. In summary, the error on the total volume is:

$$\Delta\mathcal{V} = \sqrt{\left(\frac{\partial \hat{\mathcal{V}}}{\partial \hat{c}}\Delta c\right)^2 + \left(\frac{\partial \hat{\mathcal{V}}}{\partial \hat{\beta}}\Delta\beta\right)^2}$$

or, more conservatively:

$$\Delta\mathcal{V} = \left|\frac{\partial \hat{\mathcal{V}}}{\partial \hat{c}}\right|\Delta c + \left|\frac{\partial \hat{\mathcal{V}}}{\partial \hat{\beta}}\right|\Delta\beta \tag{9}$$

5 GOOGLE TRENDS: DATA COLLECTION AND SCALING

Google Trends has several advantages over other SEO tools. First, the volumes provided are computed from the Google search engine query logs, and not from unspecified sources which may have unknown forms of bias. Second, data can be aggregated for arbitrary ranges of time and user agent languages. Most of the other tools, instead, provide monthly averages at the time of request, making it impossible to extend an experiment incrementally to new queries. Third, estimates by Google Trends are ratio-scaled, while other SEO tools provide binned values, i.e., ranges of volumes.

On the negative side, the volume provided by Google Trends is relative, not absolute. We then fix one specific query to the conventional volume of 1, and collect estimates of the volume of any other query in comparison to the specific query. Next, we scale the relative volumes to absolute volumes. In the rest of this section, we discuss some approximation introduced by relative volume calculation, and the scaling from relative to absolute volumes.

5.1 Relative volume calculation

A source of approximation in the calculation of the volume from Google Trends raw data is introduced by the computation of the ratio between the volume of a query q and the volume of the pre-fixed query f. In fact, Google Trends provides $v_f^1, \ldots, v_f^{52}$ relative volumes for f, and $v_q^1, \ldots, v_q^{52}$ relative volumes for q, namely two values for each week in our reference time period (the whole year 2017). The largest value among v_f^i's or v_q^i's is conventionally set to 100, and all the others are integers from 0 to 100 set on the basis of their fraction w.r.t. the largest value (hence, the name *relative* volume). In the following, we assume $v_f^1 = 100$ (similar reasonings apply when $v_q^1 = 100$). We aim at defining an estimator for the ratio:

$$r = \frac{\sum_{i=1}^{52} V_q^i}{\sum_{i=1}^{52} V_f^i} = \frac{\sum_{i=1}^{52} 100 \cdot V_q^i/V_f^1}{\sum_{i=1}^{52} 100 \cdot V_f^i/V_f^1}$$

where V_f^i's and V_q^i's are the true absolute volumes of f and q in the i-th week of the year, and $100 \cdot V_q^i/V_f^1$'s and $100 \cdot V_f^i/V_f^1$'s are the true relative (percentage) volumes of q and f respectively. Intuitively, r is the ratio of the total volume of q over the total volume of f. Recall that v_f^i's and v_q^i's are integer numbers. We further assume they round down[5] the true relative volume, namely $v_q^i = \lfloor 100 \cdot V_q^i/V_f^1 \rfloor$ and $v_f^i = \lfloor 100 \cdot V_f^i/V_f^1 \rfloor$. Bounds for the ratio r above are then:

$$\frac{\sum_{i=1}^{52} v_q^i}{51 + \sum_{i=1}^{52} v_f^i} \le r \le \frac{52 + \sum_{i=1}^{52} v_q^i}{\sum_{i=1}^{52} v_f^i} \tag{10}$$

An estimator for r is the central value between the bounds:

$$\hat{r} = \frac{1}{2}\left(\frac{52 + \sum_{i=1}^{52} v_q^i}{\sum_{i=1}^{52} v_f^i} + \frac{\sum_{i=1}^{52} v_q^i}{51 + \sum_{i=1}^{52} v_f^i}\right)$$

Figure 9 (left panel) shows a numerical simulation consisting of 10,000 repetitions. For each repetition, true relative volumes for 52 weeks are generated for $100 \cdot V_f^i/V_f^1$'s and $100 \cdot V_q^i/V_f^1$'s. Then, they are rounded down to obtain v_f^i's and v_q^i's respectively. Finally, the estimator $\hat{r}$ is compared to the true value r. As a parameter of the simulations, while $100 \cdot V_f^i/V_f^1$'s are uniformly generated in $[0, 100]$ (hence, the assumption $v_f^1 = 100$ is satisfied), the values $100 \cdot V_q^i/V_f^1$'s are randomly generated in $[0, 100]$ with total sum equal to a fraction of $\sum_{i=1}^{52} 100 \cdot V_f^i/V_f^1$, namely so that r is set to a desired value. In particular, Figure 9 (left panel) shows lines for $r = 0.8$ (black line), $r = 0.2$ (blue line), and $r = 0.05$ (red line). For non-small r's, the estimator $\hat{r}$ is very close[6] to the true r on average. Variance is small for larger r's. For small values of r, however, the estimator has a positive bias and very large variance. Intuitively, the weight of rounding in the bounds for r in Eq. 10 becomes considerable. Such weight can be kept low if the volumes of q and f are close to each other. For such a reason, our implementation computing the relative volume of a candidate query q consists of first comparing q with the pre-fixed query f. If their estimated ratio $\hat{r}$ is lower (resp., higher) than a given threshold, then another query f' with lower (resp., higher) relative volume is chosen such that the estimated ratio is within an expected range. The search for f' follows a binary search pattern among all queries whose relative volume has been already estimated. If no f' satisfies the condition, the estimation for q is suspended, until some other query f' will subsequently be estimated that meets the constraint. Finally, the volume of q relative to f is calculated as the product of the volume of q relative to f' multiplied by the volume of f' relative to f. Figure 9 (right panel) shows the benefits of this two-steps procedure against the single step comparing q to f. Here, $r = 0.01$ and f' is chosen with relative volume of 0.2, which is 20 times the one of q. In our actual implementation, we are even more strict. We look for an f' such that the ratio of q relative to f' is is between $4/5 = 0.8$ and $5/4 = 1.25$.

5.2 From relative to absolute volume

We tackle now the problem of transforming the relative volumes computed using Google Trends into absolute frequencies. Basically,

[4] $\zeta'(x)$ is available in R, while $\zeta^{(1,0)}(s,a)$ must be computed numerically or with other tools, e.g., Mathematica.

[5] Rounding to the closest integer is another option, which seems not consistent with the results of Google Trends. E.g., in any week series there is only one estimate of 100. Other estimated values are strictly lower.

[6] Strictly speaking, a t-test at 95% confidence level rejects the hypothesis that the mean of $\hat{r}/r$ is equal to 1. The values are however very close for the whole region of explored parameters.

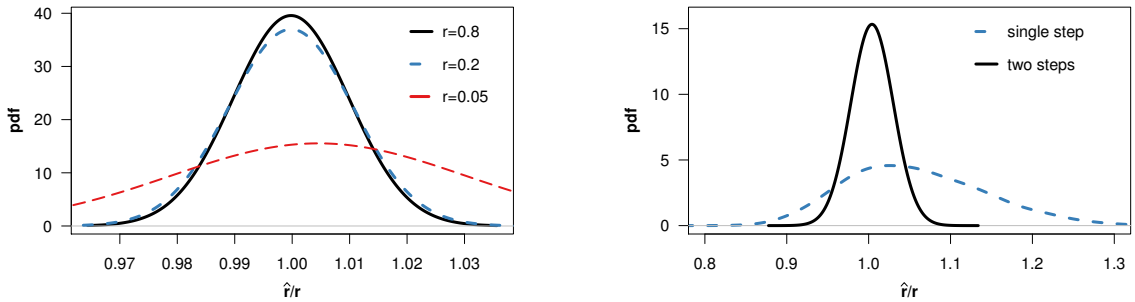

Figure 9: Simulation on estimation $\hat{r}$ of relative volume r of Google Trends. Left: at variation of r. Right: single step vs two steps procedure for $r = 0.01$.

we rely on correlating with absolute volumes provided by an external "ground truth" source. For instance, many SEO tools provide absolute estimations. Commercial tools are supposed to be more reliable than free tools, yet their fees are expensive for a large sample of queries. Moreover, in most cases such tools provide binned absolute estimations. This complicates the statistical correlation model. We consider in this sub-section the case of SEO tools with *continuous* absolute estimations.

As before we consider a sample of n queries, and for the i-th query q_i, let V_i be its true absolute volume. We cannot observe V_i, but we have two related quantities: (1) Google Trends provides a rescaled estimate $X_i = V_i/g$, but g is not known; (2) another SEO tool provides an absolute estimate of the volume of V_i, which we call Y_i. Our objective is to estimate g and therefore the absolute volume V_i. The problem is complicated by two facts: measurements computed from Google Trends are actually estimations of relative volumes (see previous subsection); and, values provided by the other SEO tool are noisy estimates of the true volume. A sensible model taking into account the two sources of noise is:

$$X_i = \frac{V_i}{g\xi_i}, \qquad Y_i = V_i\eta_i$$

where ξ_i are independent and positive error terms due to relative volume estimation[7] and are characterized by a mean value $\mathbb{E}[\xi_i] = \bar{\xi} \simeq 1$ and variance $Var[\xi_i] = z_i^2$. Similarly η_i are independent and positive random variables with mean 1 (i.e., we assume that the other SEO tool is not biased) and variance s_i^2. Note that we are not excluding the possibility that the variance of ξ_i and/or η_i depends on the rank and/or on the volume of a query – since volume of popular queries is easier to estimate. Finally, we will assume that ξ_i and η_j are mutually independent for any i and j. Combining the two expressions we obtain a relation between observable quantities:

$$Y_i = gX_i\xi_i\eta_i \tag{11}$$

Let us consider different estimators of the constant g. In the case of continuous data we compare three of them:

(1) The ratio based estimator defined as:

$$\hat{g}_1 = \frac{1}{n}\sum_{i=1}^{n}\frac{Y_i}{X_i} \tag{12}$$

[7]With the terminology of Section 5.1, ξ_i^{-1} is the rounding error, whose density is shown in Figure 9.

It is easy to show that it has the property:

$$\mathbb{E}[\hat{g}_1] = \bar{\xi}g; \qquad Var[\hat{g}_1] = \frac{g^2}{n^2}\sum_{i=1}^{n}\left(z_i^2s_i^2 + z_i^2 + \bar{\xi}^2s_i^2\right)$$

This estimator has a bias given by $\bar{\xi}$, which can be assumed very small as shown in the previous subsection. Moreover when the sample size $n \to \infty$ it is $Var[\hat{g}_1] \to 0$, i.e. the error on the estimator asymptotically vanishes.

(2) The estimator from the linear regression over the logarithms of the values:

$$\log Y_i = a + b\log X_i + \epsilon_i \tag{13}$$

and then setting $\hat{g}_2 = e^a$.

(3) The estimator from the regression:

$$\log X_i = A + B\log Y_i + \epsilon_i \tag{14}$$

given by $\hat{g}_3 = e^{-A/B}$

In the numerical simulations below we will assume that η_i follows a lognormal distribution, which is positive and has relatively large fluctuations. Given the parameters μ and σ characterizing the lognormal distribution, it is $\mathbb{E}[\eta_i] = \exp(\mu + \sigma^2/2)$ and $Var[\eta_i] = (e^{\sigma^2} - 1)(\mathbb{E}[\eta_i])^2$. In order to have $\mathbb{E}[\eta_i] = 1$, it must be $\mu = -\sigma^2/2$. Moreover we assume for simplicity that $\bar{\xi} = 1$ and $Var[\xi_i] = 0$, i.e. we neglect the approximation error in the calculation of relative volume from Google Trends data. As data generating process, we consider a Zipf's law distribution of the volumes V_i's, with parameters as in (3), and a non-uniform random sampling from it. Also, we consider two noise levels: $\sigma = 0.03$, which leads to a standard deviation of η_i equal to $s \simeq 0.03$; and $\sigma = 0.3$, which leads to $s \simeq 0.31$. Then, we fix $g = 2.75$ and estimate it using the above three estimator for 10,000 runs. Figure 10 shows the densities of the estimated g with the three methods. The ratio based estimator of Eq. 12 performs much better than the ones based on regression and this advantage is larger when the noise term has large variance.

6 EMPIRICAL ANALYSIS

We generated a sample of 120K queries by crawling popular Italian websites about recipes and cooking. The list of websites was compiled with the help of web marketing experts and by looking at

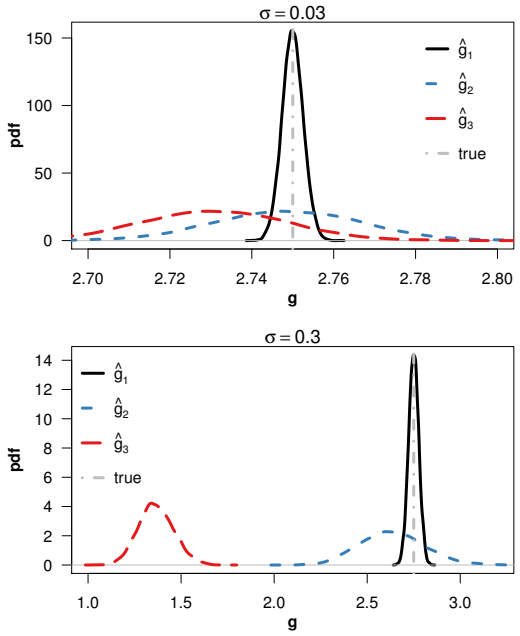

Figure 10: Density estimation of the estimated parameter g. Top plot: small noise. Bottom plot: large noise.

the rankings of SEO tools[8]. We then submitted the 120K queries to a few SEO tools to collect the estimated volume of each query for the reference year 2017 and for Italian user agents. Considering a whole year prevents seasonal bias in data. Query crawling, cleaning, and collection of Google Trends volumes took about 2 months. The process required manual inspection of crawled queries, with a few iterations to correct bugs, to support new hypotheses, etc. Even though the collection of Google Trends data was automated in a script[9], there is a daily bound on the number of invocations to the Google Trends service, which makes such a step time-consuming.

6.1 Google Trends with absolute volumes

We obtained non-zero estimates by Google Trends for about 18.5K queries out of the 120K in the sample. The resulting rank-volume distribution is shown in Figure 1. The remaining queries belong to the long tail, for which Google Trends returns a relative (hence, absolute) estimated volume of zero.

The estimators of the scaling factor g discussed in Section 5.2 require that the absolute estimates provided by an external ground truth are not biased. This assumption cannot be verified in general, e.g., SEO tools do not disclose sufficient information due to IPR restrictions. Since the bias of such tools is unknown, the choice of which one to use for estimating g relies only on the trustworthiness on one specific tool over the others.

Google Search Console[10] (GSC) is a tool that provides to website owners (a.k.a., publishers) summary statistics about the number of impressions and the ranking of the website in Search Engine Result Pages (SERPs). We considered a specific website for which

[8]E.g., https://serpstat.com
[9]We used PyTrends APIs (https://github.com/GeneralMills/pytrends).
[10]https://www.google.com/webmasters

v	$v/12$	$\hat{N}_v$	ΔN_v	$\hat{\mathcal{V}}_v$	$\Delta\mathcal{V}_v$
12	1	269,214,520	± 18,507,467	14,169.58 M	± 827.70 M
120	10	13,770,732	± 815,062	7,171.15 M	± 353.96 M
1,200	100	704,394	± 33,959	3,591.35 M	± 145.83 M
12,000	1,000	36,031	± 1,444	1,760.23 M	± 56.86 M
120,000	10,000	1,843	± 56	823.63 M	± 20.30 M
600,000	50,000	231	± 5	457.12 M	± 9.06 M

Table 1: Estimated N_v and $\mathcal{V}_v$ for queries with at least v searches in 2017. $v/12$ is the monthly average of v.

we had access to its GSC statistics. The website ranked about first in 2017 for 41 queries belonging to our sample. For such queries, the absolute volume is then equivalent to the number of impressions reported by GSC. In summary, we have ground truth volumes (or very close to it) for such set of queries. Using the estimator of Eq. 12, we found $\hat{g}_1 = 6,466.6$, that is the pre-fixed query with relative volume 1 was searched 6,466.6 times in the whole 2017 year, i.e., an average of 538.9 times per month. Figure 1 shows the rank-volume distribution where relative volume X_i has been scaled to $V_i = X_i \cdot \hat{g}_1$.

A drawback of using GSC is the low number of ground truth queries, only 41. As a second option, we consider the well-recognized SemRush[11] tool. We were able to collect volume estimates for 1,688 queries in our sample, using the paid service version of the tool. The resulting estimated scaling factor $\hat{g}_1 = 6,114$ is very close to the one obtained from GSC data.

6.2 Estimation of total query volume

Let us now apply the estimation model designed in Section 4 to the empirical data of Google Trends volumes scaled using GSC data. As shown by the red line fit in Figure 1 (left panel), the NLS regression estimates[12] are:

$$\hat{\beta} = 0.7745085 \quad \hat{c} = e^{17.5189} = 40,584,860.$$

Their statistical errors are moderately low:

$$\Delta\beta = 0.002490065 \quad \Delta c = 199,263.$$

We can now use Eq. 4 for estimating the number N_v of queries having a volume of at least v, using Eq. 8 for calculating the statistical error ΔN_v. Similarly, Eq. 7 can be used for estimating the total volume $\mathcal{V}_v$ of queries having a volume of at least v, and Eq. 9 for calculating its statistical error $\Delta\mathcal{V}_v$. Table 1 reports the estimates for a few values of v. As a means of comparison, the total empirical volume of the 18.5K queries in our sample amounts at 1,057M searches. Such a large number is consistent with the fact that the sample is not uniform, but highly ranked queries are more likely to be in the sample. Moreover, it also gives confidence that the sample is sufficiently large (as per empirical volume) to correctly estimate the true volume. According to the simulations of Section 4, the values $\hat{N}_v$ and $\hat{\mathcal{V}}_v$ may overestimate the true N_v and $\mathcal{V}_v$ respectively, if some sketchy approximation is introduced in the query volume data by Google Trends (or by any other SEO tools we might haved used in place of it). In case of noisy data,

[11]https://www.semrush.com
[12]The figure also shows the rank $max = 1725$ determined with the CSN method. The NLS fit is considered for the top max queries.

instead, under or overestimation may occur. The amount of such errors depend on the unknown amount of approximation or noise in the Google Trends data. Moreover, it is worth to stress that the reported statistical errors do not take into account such noise, but only the error of the parameter estimation procedures (assuming noiseless data).

7 CONCLUSIONS

We studied the problem of estimating the total search volume of queries belonging to a specific domain. By doing the sensible assumption that the unobserved rank distribution of absolute volumes follows a Zipf's law, our method can be decomposed in two parts. First, by comparing Google Trends data with results from SEO tools, we convert the relative volumes obtained from Google Trends for a subset of queries into absolute volumes. Second, we use the estimated absolute volumes of the subset of queries to infer the total volume of the queries of the domain. In doing this, we carefully took into account different sources of error (round-off by Google Trends and observational noise). We were also able to find the total number and the total volume of the queries in the domain which have been searched at least v times in a given time period. A large set of numerical simulations have supported the validity of our methods. Finally, we presented an empirical application to the estimation of the volume of the domain *recipes and cooking*, providing also error bars for the estimates. This kind of information is extremely useful in web marketing research and advertising.

The *first critical issue* for extending our analysis to other domains consists of checking the hypothesis that the population of queries in the domain is Zipfian. As shown in Figure 1, empirical data on the domain of recipes and cooking appears to be Zipfian. This motivated our assumption that the reference population, namely the queries searched in a reference domain, follows a Zipf's law. Ref. [9] points out that the granularity and extent of a reference population should exhibit a "coherence" property. This is particularly relevant, since splitting or merging two Zipfian sets does not necessarily yield another Zipfian set, hence the actual definition of what is and what is not in a domain is essential in meeting our assumption. The domain considered in this paper has well-defined boundaries that make it reasonably coherent.

The *second critical issue* is the construction of the sample set of queries. As shown by the numerical simulations, the capability of inferring the total volume significantly depends on the ability of selecting in the investigated sample queries which have likely high rank in the population (this is related to the parameter p in the non-uniform sampling). This set can be constructed either by resorting to domain's experts or, as we did in this paper, by crawling a set of specialized websites. Finding estimators which are (more) robust to the choice of the sample of queries is certainly an interesting potential extension of our approach to the case when it is costly or unfeasible to construct controlled samples.

The *third critical issue* is concerned with understanding which type of noise is likely to be present in empirical data provided by Google Trends or other SEO tools. In this paper, we considered three possible scenarios: uniform sampling alone, or together with normally distributed noise (noisy sampling), or together with count-min sketch like approximation (sketchy sampling). Other scenarios can be conceived, e.g., noise due to data anonymization [7, 13]. Further work is necessary to test which scenario fits better for a given SEO tool.

Finally, the *fourth critical issue* in our approach is which SEO tool to use for collecting volume of queries in the empirical sample. We relied on Google Trends, which provides relative volumes, and had to resort to GSC or other SEO tools as ground truth for determining a scaling factor. An alternative is to use SEO tools directly for collecting empirical absolute volumes. There are limitations of such tools which motivated our choice of using Google Trends – see beginning of Section 5. One of the issues is that they provide binned data. This means that the estimators of c and β might have to be reconsidered, e.g., by resorting to extensions of the CSN method to binned data [18].

ACKNOWLEDGMENTS

The work in this paper has been partially supported by a research grant by *ForTop S.R.L.* (https://www.fortop.it/en) on the topic: *Data-driven analysis of search engine query market.* We are grateful to L. Barsotti, S. Camarda, P. Ferragina, R. Guidotti, M Marino, and A. Monreale for several stimulating discussions.

REFERENCES

[1] L.A. Adamic and B.A. Huberman. 2002. Zipf's law and the Internet. *Glottometrics* 3 (2002), 143–150.
[2] R. A. Baeza-Yates, A. Gionis, F. Junqueira, V. Murdock, V. Plachouras, and F. Silvestri. 2007. The impact of caching on search engines. In *SIGIR*. ACM, 183–190.
[3] R. A. Baeza-Yates and A. Tiberi. 2007. Extracting semantic relations from query logs. In *KDD*. ACM, 76–85.
[4] A. Bookstein. 1990. Informetric distributions, part I: Unified overview. *JASIS* 41, 5 (1990), 368–375.
[5] A. Clauset, C. R. Shalizi, and M. E. J. Newman. 2009. Power-Law Distributions in Empirical Data. *SIAM Rev.* 51, 4 (2009), 661–703.
[6] G. Cormode, M. N. Garofalakis, P. J. Haas, and C. Jermaine. 2012. Synopses for Massive Data: Samples, Histograms, Wavelets, Sketches. *Foundations and Trends in Databases* 4, 1-3 (2012), 1–294.
[7] G. Cormode, T. Kulkarni, and D. Srivastava. 2018. Constrained Private Mechanisms for Count Data. In *ICDE*. IEEE, 845–856.
[8] G. Cormode and S. Muthukrishnan. 2005. An improved data stream summary: the count-min sketch and its applications. *J. Algorithms* 55, 1 (2005), 58–75.
[9] M. Cristelli, M. Batty, and L. Pietronero. 2012. There is More than a Power Law in Zipf. *Scientific Reports* 2 (2012), 812.
[10] S. Ding, J. Attenberg, R. A. Baeza-Yates, and T. Suel. 2011. Batch query processing for web search engines. In *WSDM*. ACM, 137–146.
[11] C. Gillespie. 2015. Fitting Heavy Tailed Distributions: The poweRlaw Package. *J. of Stat. Software* 64, 2 (2015), 1–16.
[12] M.L. Goldstein, S.A. Morris, and G.G. Yena. 2004. Problems with fitting to the power-law distribution. *European Physical Journal B* 41 (2004), 255–258.
[13] L. Melis, G. Danezis, and E. De Cristofaro. 2016. Efficient Private Statistics with Succinct Sketches. In *NDSS*. The Internet Society.
[14] M.E.J. Newman. 2005. Power laws, Pareto distributions and Zipf's law. *Contemporary Physics* 46, 5 (2005), 323–351.
[15] A. Orlitskya, A.T. Sureshb, and Y. Wuc. 2016. Optimal prediction of the number of unseen species. *Proceedings of the National Academy of Sciences USA* 113 (2016), 13283–13288.
[16] C. Petersen, J. Grue Simonsen, and C. Lioma. 2016. Power Law Distributions in Information Retrieval. *ACM Trans. Inf. Syst.* 34, 2 (2016), 8:1–8:37.
[17] L. Vaughan and Y. Chen. 2015. Data mining from web search queries: A comparison of Google Trends and Baidu Index. *JASIST* 66, 1 (2015), 13–22.
[18] Y. Virkar and A. Clauset. 2014. Power-Law Distributions in Binned Empirical Data. *The Annals of Applied Statistics* 8, 1 (2014), 89–119.