Rules of Thumb for Information Acquisition from Large and Redundant Data

Wolfgang Gatterbauer

Computer Science and Engineering,
University of Washington, Seattle
gatter@cs.washington.edu

Abstract. We develop an abstract model of information acquisition from redundant data. We assume a random sampling process from data which contain information with bias and are interested in the fraction of information we expect to learn as function of (*i*) the sampled fraction (*recall*) and (*ii*) varying bias of information (*redundancy distributions*). We develop two rules of thumb with varying robustness. We first show that, when information bias follows a Zipf distribution, the 80-20 rule or Pareto principle does surprisingly not hold, and we rather *expect to learn less than 40% of the information when randomly sampling 20% of the overall data*. We then analytically prove that for large data sets, randomized sampling from power-law distributions leads to "truncated distributions" with the same power-law exponent. This second rule is very robust and also holds for distributions that deviate substantially from a strict power law. We further give one particular family of power-law functions that remain completely invariant under sampling. Finally, we validate our model with two large Web data sets: link distributions to web domains and tag distributions on delicious.com.

1 Introduction

The 80-20 rule (also known as Pareto principle) states that, often in life, 20% of effort can roughly achieve 80% of the desired effects. An interesting question is as to weather this rule also holds in the context of information acquisition from redundant data. Intuitively, we know that we can find more information on a given topic by gathering a larger number of data points. However, we also know that the marginal benefit of knowing additional data decreases with the size of the sampled corpus. Does the 80-20 rule hold for information acquisition from redundant data? Can we learn 80% of URLs on the Web by parsing only 20% of the web pages? Can we learn 80% of the used vocabulary by looking at only 20% of the tags? Can we learn 80% of the news by reading 20% of the newspapers? More generally, *can we learn 80% of all available information* in a corpus *by randomly sampling 20% of data* without replacement?

In this paper, we show that when assuming a Zipf redundancy distribution, the Pareto principle does not hold. Instead, we rather expect to see less than 40% of the available information. To show this in a principled, yet abstract

P. Clough et al. (Eds.): ECIR 2011, LNCS 6611, pp. 479–490, 2011.

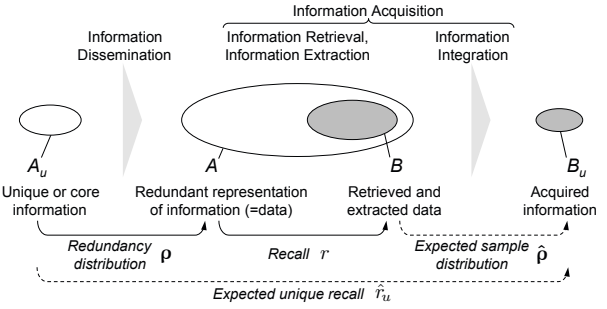


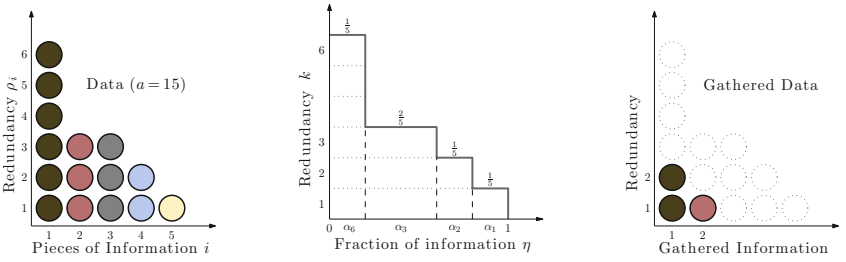
Fig. 1. Processes of information dissemination and information acquisition. We want to predict the fraction of information we can learn ($\hat{r}_u$) as a function of recall (r) and the bias in the data (redundancy distribution $\boldsymbol{\rho}$).

fashion, we develop an *analytic sampling model* of information acquisition from redundant data. We assume the *dissemination* of relevant information is biased, i.e. different pieces of information are more or less frequently represented in available sources. We refer to this bias as *redundancy distribution* in accordance with work on redundancy in information extraction [8]. Information acquisition, in turn, can be conceptually broken down into the subsequent steps of IR, IE, and II, i.e. visiting a fraction r of the available sources, extracting the information, and combining it into a unified view (see Fig. 1). Our model relies on only three simple abstractions: (1) we consider a purely *randomized sampling* process without replacement; (2) we do not model *disambiguation* of the data, which is a major topic in information extraction, but not our focus; and (3) we consider the process in the limit of *infinitely large data sets*. With these three assumptions, we estimate the success of information acquisition as function of the (*i*) *recall* of the retrieval process and (*ii*) *bias in redundancy* of the underlying data.

Main contributions. We develop an analytic model for the information acquisition from redundant data and (1) derive the 40-20 rule, a modification of the Pareto principle which has not been stated before. (2) While power laws do not remain invariant under sampling in general [19], we prove that one particular power law family does remain invariant. (3) While other power laws do not remain invariant in their overall shape, we further prove that the "core" of such a frequency distribution does remain invariant; this observations allows us to develop a second rule of thumb. (4) We validate our predictions by randomly sampling from two very large real-world data sets with power-law behavior. All proofs and more details are available in the full version of this paper [13].

2 Basic Notions Used Throughout This Paper

We use the term *redundancy* as synonym for frequency or multiplicity. We do so to remain consistent with the term commonly used in web information extraction, referring to the redundant nature of information on the Web. The notions of data and information are defined in various and partly contradicting ways



(a) $\boldsymbol{\rho}=(6,3,3,2,1)$ (b) $\boldsymbol{\alpha}=(\frac{1}{5},\frac{1}{5},\frac{2}{5},0,0,\frac{1}{5})$ (c) $r=\frac{3}{15}, r_u=\frac{2}{5}$

Fig. 2. (a): Redundancy distribution $\boldsymbol{\rho}$. (b): Representation as *redundancy frequency distribution* $\boldsymbol{\alpha}$. (c): Recall $r=\frac{3}{15}$ and *unique recall* $r_u=\frac{2}{5}$.

in the information retrieval, information extraction, database and data integration literature. In general, their difference is attributed to novelty, relevance, organization, available context or interpretation. The most commonly found understanding is that of *data as representation of information* which can become information when it is interpreted as new and relevant in a given context [4]. In this work, we follow this understanding of data as "raw" information and use the term data for the partly redundant representation of information.

Let a be the total number of data items and a_u the number of unique pieces of information among them. *Average redundancy* ρ is simply their ratio $\rho=\frac{a}{a_u}$. Let ρ_i refer to the redundancy of the i-th most frequent piece of information. The *redundancy distribution* $\boldsymbol{\rho}$ (also known as rank-frequency distribution) is the vector $\boldsymbol{\rho}=(\rho_1,\ldots,\rho_{a_u})$. Figure 2a provides the intuition with a simple balls-and-urn model: Here, each color represents a piece of information and each ball represents a data item. As there are 3 red balls, redundancy of the information "color = red" is 3. Next, let α_k be the fraction of information with redundancy equal to k, $k \in [k_{\max}]$. A *redundancy frequency distribution* (also known as count-frequency plot) is the vector $\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_{k_{\max}})$. It allows us to describe redundancy without regard to the overall number of data items a (see Fig. 2b) and, as we see later, an analytic treatment of sampling for the limit of infinitely large data sets. We further use the term *redundancy layer* (also known as complementary cumulative frequency distribution or ccfd) η_k to describe the fraction of information that appears with redundancy $\geq k$: $\eta_k=\sum_{i=k}^{k_{\max}}\alpha_i$. For example, in Fig. 2a and Fig. 2b, the fraction of information with redundancy at least 3 is $\eta_3=\alpha_3+\alpha_6=\frac{2}{5}+\frac{1}{5}=\frac{3}{5}$. Finally, *recall* is the well known measure for the coverage of a data gathering or selection process. Let b be a retrieved subset of the a total data items. Recall is then $r=\frac{b}{a}$. We define *unique recall* as is its counterpart for unique data items. Thus, it measures the coverage of information. Let b_u be the number of unique pieces of information among b, and a_u the number of unique pieces of information among a. Unique recall r_u is then $r_u=\frac{b_u}{a_u}$. We illustrate again with the urns model: assume that we randomly gather 3 from the 15 total balls (recall $r=\frac{3}{15}$) and that, thereby, we learn 2 colors out of the 5 total available colors (Fig. 2c). Unique recall is thus $r_u=\frac{2}{5}$ and the sample redundancy distribution is $\hat{\boldsymbol{\rho}}=(2,1)$.

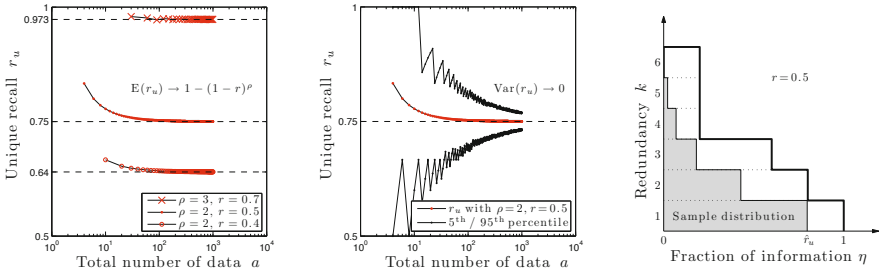


Fig. 3. (a, b): Random sampling from an urn filled with a balls in a_u different colors. Each color appears on exactly $\rho = \frac{a}{a_u}$ balls. (c): Normalized sample distribution in grey with unique recall $\hat{r}_u \approx 0.8$ for $r = 0.5$ from $\boldsymbol{\alpha}$ in Fig. 2b.

3 Unique Recall

We next give an analytic description of *sampling without replacement* as function of recall and the bias of available information in the limit of very large data sets.

Proposition 1 (Unique recall $\hat{r}_u$). *Assume randomized sampling without replacement with recall $r \in [0,1]$ from a data set with redundancy frequency distribution $\boldsymbol{\alpha}$. The expected value of unique recall for large data sets is asymptotically concentrated around* $\boxed{\hat{r}_u = 1 - \sum_{k=1}^{k_{\max}} \alpha_k \, (1-r)^k}$.

The proof applies Stirling's formula and a number of analytic transformations to a combinatorial formulation of a balls-and-urn model. The important consequence of Prop. 1 is now that unique recall can be investigated without knowledge of the actual number of data items a, but by just analyzing the normalized redundancy distributions. Hence, we can draw general conclusions for families of redundancy distributions assuming very large data sets. To simplify the presentation and to remind us of this limit consideration, we will use the hat symbol and write $\hat{r}_u$ for $\lim_{a\to\infty} \mathbf{E}\,(r_u) \simeq \mathbf{E}\,(r_u)$.

Figure 3 illustrates this limit value with two examples. First, assume an urn filled with a balls in a_u different colors. Each color appears on exactly two balls, hence $\rho = 2$ and $a = 2a_u$. Then the expected value of unique recall r_u (fraction of colors sampled) is converging towards $1-(1-r)^\rho$ and its variance towards 0 for increasing numbers of balls a (Fig. 3a, Fig. 3b). For example, keeping $\rho = 2$ and $r = 0.5$ fixed, and varying only $a = 4, 6, 8, 10, \ldots$, then unique recall varies as $r_u = 0.83, 0.80, 0.79, 0.78, \ldots$, and converges towards $\hat{r}_u = 0.75$. At $a = 1000$, r_u is already 0.7503 ± 0.02 with 90% confidence. Second, assume that we sample 50% of balls from the distribution $\boldsymbol{\alpha} = (\frac{1}{5}, \frac{1}{5}, \frac{2}{5}, 0, 0, \frac{1}{5})$ of Fig. 2b. Then we can expect to learn $\approx$80% of the colors if a is very large (Fig. 3c). In contrast, exact calculations show that if $a = 15$ as in Fig. 2a, then the actual value of $\mathbf{E}\,(r_u)$ is around $\approx$79% or $\approx$84% for sampling 7 or 8 balls, respectively (note that 7.5 is 50% of 15). Thus, Prop. 1 calculates the exact asymptotic value only for the limit, but already gives very good approximations for large data sets.

4 Unique Recall for Power Law Redundancy Distributions

Due to their well-known ubiquity, we will next study *power law redundancy distributions*. We distinguish three alternative definitions: (1) power laws in the redundancy distributions, (2) in the redundancy frequencies, and (3) in the redundancy layers. These three power laws are commonly considered to be different expressions of the identical distribution [2,16] because they have the same tail distribution[1], and they are in fact identical in a continuous regime. However, for discrete values, these three definitions of power laws actually produce different distributions and have different unique recall functions. We will show this next.

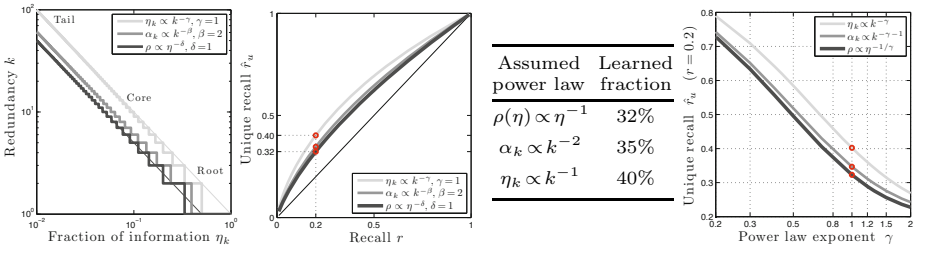
Power laws in the redundancy distribution $\boldsymbol{\rho}$. This distribution arises when the frequency or redundancy ρ of an item is proportional to a power law with exponent δ of its rank i: $\rho(i) \propto i^{-\delta}$, $i \in [a_u]$. Two often cited examples of such power law redundancy distributions where $\delta \approx 1$ are the frequency-rank distribution of words appearing in arbitrary corpora and the size distribution of the biggest cities for most countries. These are called "Zipf Distribution" after [20]. Using Prop. 1 we can derive in a few steps $\hat{r}_{u\rho}(r,\delta) = 1 - \sum_{k=1}^{\infty} \left((2k-1)^{-\frac{1}{\delta}} - (2k+1)^{-\frac{1}{\delta}}\right)(1-r)^k$. For the particularly interesting case of $\delta = 1$, this infinite sum can be reduced to $\hat{r}_{u\rho}(r,\delta{=}1) = \frac{r}{\sqrt{1-r}} \operatorname{artanh}(\sqrt{1-r})$.

Power laws in the redundancy frequency distribution $\boldsymbol{\alpha}$. This distribution arises when a fraction of information α_k that appears exactly k times follows a power law $\alpha_k = C \cdot k^{-\beta}$, $k \in \mathbb{N}_1$. Again, using Prop. 1 we can derive in a few steps $\hat{r}_{u\alpha}(r,\beta) = 1 - \frac{\mathrm{Li}_\beta(1-r)}{\zeta(\beta)}$, where $\mathrm{Li}_\beta(x)$ is the polylogarithm $\mathrm{Li}_\beta(x) = \sum_{k=1}^{\infty} k^{-\beta} x^k$, and $\zeta(\beta)$ the Riemann zeta function $\zeta(\beta) = \sum_{k=1}^{\infty} k^{-\beta}$.

Power laws in the redundancy layers $\boldsymbol{\eta}$. This distribution arises when the redundancy layers $\eta_k \in [0,1]$ follow a power law $\eta_k \propto k^{-\gamma}$. From $\eta_1 = 1$, we get $\eta_k = k^{-\gamma}$ and, hence, $\alpha_k = k^{-\gamma} - (k+1)^{-\gamma}$. Using again Prop. 1, we get in a few steps $\hat{r}_{u,\eta}(r,\gamma) = \frac{r}{1-r} \mathrm{Li}_\gamma(1-r)$. For the special case of $\gamma = 1$, we can use the property $\mathrm{Li}_1(x) = -\ln(1-x)$ and simplify to $\hat{r}_{u,\eta}(r,\gamma{=}1) = -\frac{r \ln r}{1-r}$.

Comparing unique recall for power laws. All three power laws show the typical *power law tail* in the loglog plot of the redundancy distribution (loglog rank-frequency plot), and it is easily shown that the exponents can be calculated from each other according to Fig. 4e However, the distributions are actually different at the *power law root* (Fig. 4a) and also lead to different unique recall functions. Figure 4b shows their different unique recall functions for the particular power law exponent of $\delta{=}1$ ($\gamma{=}1$, $\beta{=}2$), which is assumed to be the most common redundancy distribution of words in a large corpus [20] and many other frequency distributions [16,17]. Given our normalized framework, we can now ask the interesting question: *Does the 80-20 rule hold for information acquisition assuming a Zipf distribution?* Put differently, if we sample 20% of the total

[1] With *tail of a distribution*, we refer to the part of a redundancy distribution for $\eta \to 0$, with *root* to $\eta \to 1$, and with *core* to the interval in between (see Fig. 4a).



(a) (b)

Assumed power law	Learned fraction
$\rho(\eta) \propto \eta^{-1}$	32%
$\alpha_k \propto k^{-2}$	35%
$\eta_k \propto k^{-1}$	40%

(c)

(d)

Power law	Exp	δ	β	γ	Defined range	Example values				
$\boldsymbol{\rho}$: $\rho(\eta) \propto \eta^{-\delta}$	$\delta =$	–	$\frac{1}{\beta-1}$	$\frac{1}{\gamma}$	$0 < \delta < \infty$	5	2	1.3	1	0.5
$\boldsymbol{\alpha}$: $\alpha_k \propto k^{-\beta}$	$\beta =$	$\frac{1}{\delta}+1$	–	$\gamma+1$	$1 < \beta < \infty$	1.2	1.5	1.77	2	3
$\boldsymbol{\eta}$: $\eta_k \propto k^{-\gamma}$	$\gamma =$	$\frac{1}{\delta}$	$\beta-1$	–	$0 < \gamma < \infty$	0.2	0.5	0.77	1	2

(e)

Fig. 4. The three power law redundancy distributions (e) have the same *power law tail* and *power law core* but different *power law roots* in the loglog redundancy plot (a). This leads to different unique recall functions (b&d), and different fractions of information learned after sampling 20% data (c).

amount of data (e.g., read 20% of a text corpus, or look at 20% of all existing tags on the Web), what percentage of the contained information (e.g., fraction of different words in a corpus or the tagging data) can we expect to learn if redundancy follows a Zipf distribution? Figure 4c lists the results for the three power law distributions and shows that we can only expect to learn between 32% and 40% of the information, depending on which from the three definitions we choose. Note that we can apply this rule of thumb without knowing the total amount of available information. Also note that these numbers are sensitive to the power law root and, hence, to deviations from an ideal power law. This is also why unique recall diverges for our 3 variations of power law definitions in the first place (Fig. 4a). Finally, Fig. 4d shows that the power law exponent would have to be considerably different from $\gamma = 1$ to give a 80-20 rule.

Rule of thumb 1 (40-20 rule). *When randomly sampling 20% of data whose redundancy distribution follows an exact Zipf distribution, we expect to learn less than 40% of the contained information.*

5 K-Recall and the Evolution of Redundancy Distributions

So far, we were interested in the expected fraction r_u of information we learn when we randomly sample a fraction r of the total data. We now generalize the question and derive an analytic description of the overall shape of the *expected sample redundancy distribution* (Fig. 3c). As it turns out, and what will become clear in this and the following section, the natural way to study and solve this question is again to *analyze the horizontal "evolution" of the redundancy layers* $\boldsymbol{\eta}$

during sampling. To generalize unique recall r_u, we define k-recall as the fraction r_{uk} of information that has redundancy $\geq k$ and also appears at least k times in our sample. More formally, let a_{uk} be the number of unique pieces of information with redundancy $\geq k$ in a data set, and let b_{uk} be the number of unique pieces of information with redundancy $\geq k$ in a sample. *K-recall* r_{uk} is then the fraction of a_{uk} that has been sampled: $r_{uk} = \frac{b_{uk}}{a_{uk}}$. The special case r_{u1} is then simply the so far discussed unique recall r_u. We assume large data sets throughout this and all following section without always explicitly using the hat notation $\hat{r}_{uk}$.

K-recall has its special relevance when sampling from partly unreliable data. In such circumstances, the general fall-back option is to assume a piece of information to be true when it is independently learned from at least k different sources. This approach is used in statistical polling, in many artificial intelligence applications of learning from unreliable information, and in consensus-driven decision systems: Counting the number of times a piece of information is occurring (its *support*) is used as strong indicator for its truth. As such, *to believe a piece of information only when it appears at least k times in a random sample* serves as starting point from which more complicated polling schemes can be conceived. In this context, r_{uk} gives the ratio of information that we learn *and* consider true (it appears $\geq k$ times in our sample) to the overall information that we would consider true if known to us (it appears $\geq k$ times in the data set) (Fig. 5a).

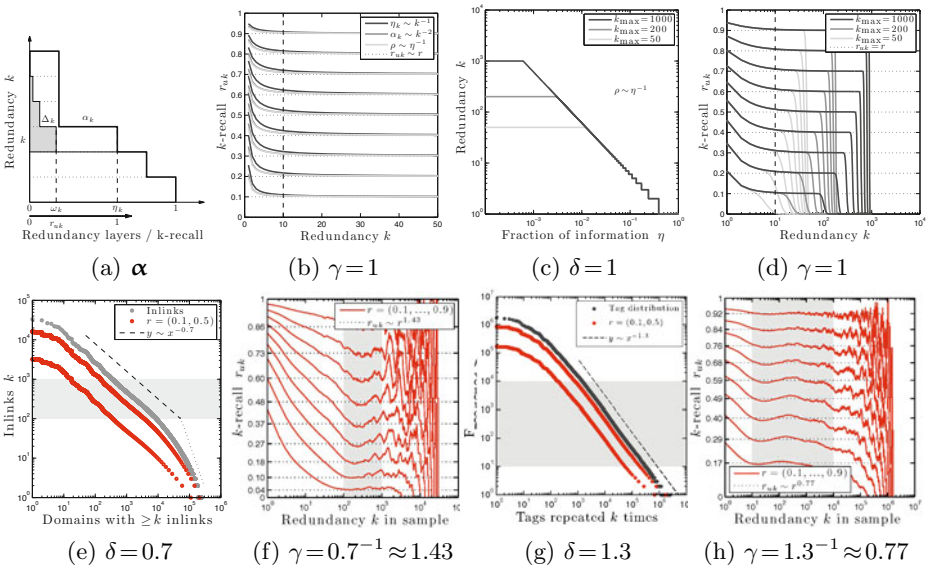
We also introduce a variable ω_k for the fraction of total information we get in our sample that appears at least k times instead of just once. Note that $\omega_k = \eta_k r_{uk} = \frac{b_{uk}}{a_u}$. All ω_k with $k \in [k_{\max}]$ together form the vector $\boldsymbol{\omega}$ representing the *sample redundancy layers* in a random sample with $r \in [0, 1]$. As r increases from 0 to 1, it "evolves" from the $k_{\max}$-dimensional null vector $\mathbf{0}$ to the redundancy layers $\boldsymbol{\eta}$ of the original redundancy distribution. Because of this intuitive interpretation, we call *evolution of redundancy* the transformation of a redundancy frequency distribution given by the redundancy layers $\boldsymbol{\eta}$ to the expected distribution $\boldsymbol{\omega}$ as a function of r: $\boldsymbol{\eta} \xrightarrow{r} \boldsymbol{\omega}, r \in [0, 1]$. We further use Δ_k to describe the fraction of information with redundancy exactly k: $\Delta_k = \omega_k - \omega_{k+1}$. To define this equation for all $k \in \mathbb{N}_0$, we make the convention $\omega_0 = 1$ and $\omega_k = 0$ for $k > k_{\max}$. We can then derive the following analytic description:

Proposition 2 (Sample distribution $\boldsymbol{\omega}$). *The asymptotic expectation of the fraction of information ω_k that appears with redundancy $\geq k$ in a randomly sampled fraction r without replacement from a data set with redundancy distribution $\boldsymbol{\alpha}$ is* $\boxed{\hat{\omega}_k = 1 - \sum_{y=0}^{k-1} \sum_{x=y}^{\infty} \alpha_x \binom{x}{y} r^y (1-r)^{x-y}}$ *for* $\lim_{a \to \infty}$.

The first part of the proof constructs a geometric model of sampling from infinitely large data sets with homogenous redundancy and derives the binomial distribution as evolution of the redundancy layers. The second part then applies this result to stratified sampling from arbitrary redundancy distributions.

6 The Evolution of Power Laws

Given the complexity of Prop. 2, it seems at first sight that we have not achieved much. As it turns out, however, this equation hides a beautiful simplicity for



(a) $\boldsymbol{\alpha}$ (b) $\gamma=1$ (c) $\delta=1$ (d) $\gamma=1$

(e) $\delta=0.7$ (f) $\gamma=0.7^{-1}\approx1.43$ (g) $\delta=1.3$ (h) $\gamma=1.3^{-1}\approx0.77$

Fig. 5. Given a redundancy frequency distribution $\boldsymbol{\alpha}$ and recall r. K-recall r_{uk} describes the fraction of information appearing $\geq k$ times that also appears $\geq k$ times in our sample: $r_{uk}=\frac{\omega_k}{\eta_k}$ (a). Sampling from *completely developed power laws* leads to sample distributions with the same power law tail, and $r_{uk}\approx r^\gamma$ holds independent of k for $k\gtrsim 10$ (b). *Truncated power laws* are cut off at some maximum value $k_{\max}$ (c). As a consequence, the tails of the sample distributions "break in" for increasingly lower recalls (d). However, the invariant power law core with $r_{uk}\approx r^\gamma$ is still visible. Original and sample web link distribution (e). Resulting k-recalls (f). Original and sample tag distribution on Delicious (g). Resulting k-recalls (h).

power laws: namely, their overall shape remains "almost" invariant during sampling. We will first formalize this notion, then prove it, and finally use it for another, very robust rule of thumb.

We say a redundancy distribution $\boldsymbol{\alpha}$ is *invariant under sampling* if, independent of r, the expected *normalized sample distribution* $\boldsymbol{\Delta}/\omega_1$ *is the same as the original distribution*: $\frac{\Delta_k}{\omega_1}=\alpha_k$. Hence, for an invariant distribution it holds that $\omega_k=\sum_{x=k+1}^{\infty}\Delta_x=\omega_1\sum_{x=k+1}^{\infty}\alpha_x=\omega_1\eta_k$, and, hence, r_{uk} is independent of k: $r_{uk}=\omega_1$. With this background, we can state the following lemma:

Lemma 1 (Invariant family). *The following family of redundancy distributions is invariant under sampling:* $\alpha_k=(-1)^{k-1}\binom{\tau}{k}$, *with* $0<\tau\leq 1$.

The proof of Lemma 1 succeeds by applying Prop. 2 to the invariant family and deriving $r_{uk}=r^\tau$ after application of several binomial identities. Note that the invariant family has a power law tail. We see that by calculating its asymptotic behavior with the help of the asymptotic of the binomial coefficient $\binom{\tau}{k}=\mathcal{O}\left(\frac{1}{k^{1+\tau}}\right)$, as $k\to\infty$, for $\tau\notin\mathbb{N}$. Therefore, we also have $\alpha_k=\mathcal{O}\left(k^{-(1+\tau)}\right)$ for $k\to\infty$. Comparing this equation with the power-law in the redundancy

frequency plot, $\alpha_k \propto k^{-\beta}$, we get the power-law equivalent exponent as $\beta = \tau + 1$, with $1 < \beta \leq 2$. Also note that the invariant family is not "*reasonable*" according to the definition of [1], since the mean redundancy $\sum_{k=1}^{\infty} \eta_k$ is not finite.

We next analyze sampling from *completely developed power laws*, i.e. distributions that have infinite layers of redundancy ($k_{\max} \to \infty$). Clearly, those cannot exist in real discrete data sets, but their formal treatment allows us to also consider sampling from *truncated power laws*. The latter are real-world power law distributions which are truncated at $k_{\max} \in \mathbb{N}$ (Fig. 5c). We prove that the power law core remains invariant for truncated power laws, and they, hence, appear as "almost" invariant over a large range, i.e. except for their tail and their root.

Lemma 2 (Completely developed power laws). *Randomized sampling without replacement from redundancy distributions with completely developed power law tails $\boldsymbol{\alpha}_C$ leads to sample distributions with the same power law tails.*

Theorem 1 (Truncated power laws). *Randomized sampling without replacement from redundancy distributions with truncated power law tails $\boldsymbol{\alpha}_T$ leads to distributions with the same power law core but further truncated power law tails.*

The proof for Lemma 2 succeeds in a number of steps by showing that $\lim_{i,j \to \infty} \frac{r_{ui}}{r_{uj}} = 1$ for distributions with $\boldsymbol{\alpha}_C$. The proof of Theorem 1 builds upon this lemma and shows that $\lim_{k \ll k_{\max}} \Delta_k(\boldsymbol{\alpha}_T, r) = \Delta_k(\boldsymbol{\alpha}_C, r)$. In other words, Theorem 1 states that sampling *from real-world power law distributions* leads to distributions with the same power law core but possibly different tail and root. More formally, $r_{uk} \approx r^{\gamma}$ for $k_1 < k < k_2$, where k_1 and k_2 depend on the actual distribution, maximum redundancy and the power law exponent. Both, tail and root, are usually ignored when judging whether a distribution follows a power law (cf. Figure 3 in [6]), and to the best of our knowledge, this result is new. Furthermore, it is only recently that Stumpf et al. [19] have shown that sampling from power laws does not lead to power laws in the sample, in general. Our results clarifies this result and shows that *only their tails and roots are subject to change*. Figure 5b, Fig. 5c and Fig. 5d illustrate our result.

Rule of thumb 2 (Power law cores). *When randomly sampling from a power law redundancy distribution, we can expect the sample distribution to be a power law with the same power law exponent in the core:* $\boxed{r_{uk} \approx r^{\gamma} \text{ for } k_1 < k < k_2}$.

7 Large Real-World Data Sets

Data sets. We use two large real-world data sets that exhibit power-law characteristics to verify and illustrate our rules of thumb: the number of links to web domains and the keyword distributions in social tagging applications.

(1) The first data set is a snapshot of a top level domain in the World Wide Web. It is the result of a complete crawl of the Web and several years old. The set contains 267,415 domains with 5.422,730 links pointing between them. From Fig. 5e, we see that the redundancy distribution follows a power law with

exponent $\delta = 0.7$ ($\gamma \approx 1.43$, $\beta \approx 2.43$) for $k \gtrsim 100$. Below 100, however, the distribution considerably diverges from this exponent, which is why we expect that rule of thumb 1 does not apply well. We now assume random sampling amongst all links in this data set (e.g. we randomly choose links and discover new domains) and ask: (*i*) what is the expected number of domains and their relative support (as indicated by linking to it) that we learn as function of the percentage of links seen? (*ii*) what is the fraction of domains with support $\geq k$ in the original data that we learn with the same redundancy?

(2) The second data set concerns different keywords and their frequencies used on the social bookmarking web service Delicious (`http://delicious.com`). A total number of $\approx$ 140 Mio tags are recorded of which $\approx$ 2.5 Mio keywords are distinct [5]. The redundancy distribution (Fig. 5g) follows a power law with exponent $\delta = 1.3$ ($\gamma \approx 0.77$, $\beta \approx 1.77$) very well except for the tail and the very root. Here we assume random sampling amongst all individual tags given by users (e.g. we do not have access to the database of Delicious, but rather crawl the website) and ask: (*i*) what is the expected number of different tags and their relative redundancies that we learn as function of the percentage of all tags seen? (*ii*) what is the fraction of important tags in the sample (tags with redundancy at least k) that we can also identify as important by sampling a fraction r?

Results. From Fig. 5f and Fig. 5h we see that after sampling 20% of links and tags respectively, we learn 60% and 40% of the domains and words, respectively. Hence, our first rule of thumb works well only for the second data set which better follows a power law (compare also with the prediction for arbitrary exponents in Fig. 4d). Our second rule of thumb, however, works well for both data sets: In Fig. 5f, we see that, in accordance with our predication, the horizontal lines for $r_{uk} = r^{\gamma}$ become apparent for $10^2 < k < 10^3$, and in Fig. 5h, for $10^1 < k < 10^4$ (compare with our prediction in Fig. 5d).

8 Related Work

Whereas the influence of redundancy of a search process has been widely analyzed [3,15,18], and randomized sampling used in other papers in this field [8,15], our approach is new in the way that we analytically characterize the behavior of the sampling process as a function of (*i*) the bias in redundancy of the data and (*ii*) recall of the used retrieval process. In particular, this approach allows us to prove a to date unknown characteristics of power laws during sampling. Achlioptas et al. [1] give a mathematical model that shows that traceroute sampling from Poisson-distributed random graphs leads to power laws. Their analysis is limited to "reasonable" power laws, which are such for which $\alpha > 2$ and also assumes a very concrete sampling process tailored to their context. This is different from our result which proves that completely developed power law functions retain their power law tail, and truncated power laws at least their power law core during sampling. Haas et al. [14] investigate ways to estimate the number of different attribute values in a given database. This problem is related in its background, but different in its focus. We estimate the number

of unique attributes seen after sampling a fraction and later the overall sample distribution. Stump et al. [19] show that, in general, power laws do not remain invariant under sampling. In this paper, we could show that – while not in their entirely – at least the *core of power laws remains invariant under sampling*. General balls-and-urn models have been treated in detail by Gardy [11]. Gardy showed a general theorem which contains Prop. 1 as a special case. However, she does neither investigate the behavior of power laws during sampling, nor extends this result to the evolution of the overall distribution (Prop. 2). Only the latter allowed us to investigate the overall shape of redundancy distributions during sampling. Flajolet and Sedgewick [9] study the evolution of balanced, single urn models of finite dimensions under random sampling, where dimensionality refers to the number of colors. They mainly focus on urn models of dimension 2 (i.e., balls can be of either of two colors), and also solve some special cases for higher dimensions. They further note, that there is no hope to obtain general solutions for higher dimensions, however, that special cases warrant further investigation. Using a slightly different nomenclature, we also studied a special case of balanced, single urn models, however with infinite dimension (i.e., infinite number of colors), and showed that this case allows closed analytic solutions.

In [12], we gave Prop. 1 and motivated the role of different families of redundancy distributions on the effectiveness of information acquisition. However, we did not treat the case of power laws, nor the evolution of distributions during sampling (Prop. 2). To the best of our knowledge, the main results in this paper are new. Our analytic treatment of power laws during sampling, the invariant family, and the proof that sections of power laws remain invariant are not mentioned in any prior work we are aware of (cf. [7,9,10,11,16,17]).

9 Discussion and Outlook

Our target with this paper was to develop a general model of the information acquisition process (retrieval, extraction and integration) that allows us to estimate the overall success rate when acquiring information from redundant data. With our model, we derive the 40-20 rule of thumb, an adaptation of the Pareto principle. This is a negative result as to what can be achieved, in general. A crucial idea underlying our mathematical treatment of sampling was adopting a horizontal perspective of sampling and *thinking in layers of redundancy* ("k-recall"). Whereas our approach assumes an infinite amount of data, we have shown our approximation holds very well for large data sets (see Fig. 2b). We have focused on power laws, as they are the dominant form of biased frequency distributions. Whereas Stump et al. [19] have shown that, in general, power laws do not remain invariant under sampling, we have shown that (*i*) there exists one concrete family of power laws which *does* remain invariant, and (*ii*) while power laws do not remain invariant in their tails and root, their core does remain invariant. And we have used this observation to develop a second rule of thumb which turns out to be very robust (cp. Fig. 5d with Fig. 5h).

Acknowledgements. This work was partially supported by a DOC scholarship from the Austrian Academy of Sciences, by the FIT-IT program of the Austrian Federal Ministry for Transport, Innovation and Technology, and by NSF IIS-0915054. The author would like to thank one of the major search engines for the web link data, Ciro Cattuto and the TAGora project for the Delicious tag data, and Bernhard Gittenberger and the anonymous reviewers for helpful comments. More details are available on the project page: `http://uniquerecall.com`

References

1. Achlioptas, D., Clauset, A., Kempe, D., Moore, C.: On the bias of traceroute sampling: or, power-law degree distributions in regular graphs. In: STOC, pp. 694–703 (2005)
2. Adamic, L.A.: Zipf, power-law, pareto – a ranking tutorial. Technical report, Information Dynamics Lab, HP Labs, Palo Alto, CA 94304 (October 2000)
3. Bernstein, Y., Zobel, J.: Redundant documents and search effectiveness. In: CIKM, pp. 736–743 (2005)
4. Capurro, R., Hjørland, B.: The concept of information. Annual Review of Information Science and Technology 37(1), 343–411 (2003)
5. Cattuto, C., Loreto, V., Pietronero, L.: Semiotic dynamics and collaborative tagging. PNAS 104(5), 1461–1464 (2007)
6. Chaudhuri, S., Church, K.W., König, A.C., Sui, L.: Heavy-tailed distributions and multi-keyword queries. In: SIGIR, pp. 663–670 (2007)
7. Clauset, A., Shalizi, C.R., Newman, M.: Power-law distributions in empirical data. SIAM Review 51(4), 661–703 (2009)
8. Downey, D., Etzioni, O., Soderland, S.: A probabilistic model of redundancy in information extraction. In: IJCAI, pp. 1034–1041 (2005)
9. Flajolet, P., Dumas, P., Puyhaubert, V.: Some exactly solvable models of urn process theory. Discrete Math. & Theoret. Comput. Sci. AG, 59–118 (2006)
10. Flajolet, P., Sedgewick, R.: Analytic combinatorics. CUP (2009)
11. Gardy, D.: Normal limiting distributions for projection and semijoin sizes. SIAM Journal on Discrete Mathematics 5(2), 219–248 (1992)
12. Gatterbauer, W.: Estimating Required Recall for Successful Knowledge Acquisition from the Web. In: WWW, pp. 969–970 (2006)
13. Gatterbauer, W.: Rules of thumb for information acquisition from large and redundant data. CoRR abs/1012.3502 (2010)
14. Haas, P.J., Naughton, J.F., Seshadri, S., Stokes, L.: Sampling-based estimation of the number of distinct values of an attribute. In: VLDB, pp. 311–322 (1995)
15. Ipeirotis, P.G., Agichtein, E., Jain, P., Gravano, L.: To search or to crawl? towards a query optimizer for text-centric tasks. In: SIGMOD, pp. 265–276 (2006)
16. Mitzenmacher, M.: A brief history of generative models for power law and lognormal distributions. Internet Mathematics 1(2), 226–251 (2004)
17. Newman, M.E.: Power laws, pareto distributions and zipf's law. Contemporary Physics 46(5), 323–351 (2005)
18. Soboroff, I., Harman, D.: Overview of the trec 2003 novelty track. In: TREC 2003. NIST, pp. 38–53 (2003)
19. Stumpf, M.P.H., Wiuf, C., May, R.M.: Subnets of scale-free networks are not scale-free: sampling properties of networks. PNAS 102(12), 4221–4224 (2005)
20. Zipf, G.K.: Human Behaviour and the Principle of Least Effort: an Introduction to Human Ecology. Addison-Wesley, Reading (1949)