Using Sketches to Estimate Two-way and Multi-way Associations

Ping Li*
Stanford University

Kenneth W. Church†
Microsoft Research

We should not have to look at the entire corpus (e.g., the Web) to know if two (or more) words are associated or not.[1] *A powerful sampling technique called Sketches was originally introduced to remove duplicate Web pages. We generalize sketches to estimate contingency tables and associations, using a maximum likelihood estimator to find the most likely contingency table given the sample, the margins (document frequencies) and the size of the collection. The proposed method has smaller errors and more flexibility than the original sketch method.*

Not unsurprisingly, computational work and statistical accuracy (variance or errors) depend on sampling rate, as will be shown both theoretically and empirically. Sampling methods become more and more important with larger and larger collections. At Web scale, sampling rates as low as 10^{-4} *may suffice.*

1 Introduction

Word associations (co-occurrences, or joint frequencies) have a wide range of applications including: Speech Recognition, Optical Character Recognition and Information Retrieval (IR) (Church and Hanks, 1991; Dunning, 1993; Salton, 1989; Manning and Schutze, 1999; Baeza-Yates and Ribeiro-Neto, 1999). It is easy to compute association scores for a small corpus, but more challenging to compute lots of scores for lots of data (e.g., the Web), with billions of web pages (D) and millions of word types (N). For a small corpus, one could compute pair-wise (two-way) associations by multiplying the (0/1) term-by-document matrix with its transpose (Deerwester et al., 1999). But this is probably infeasible at Web scale. Furthermore, the computation and storage cost increases exponentially for multi-way associations.

Web search engines produce estimates of page hits, as illustrated in Tables 1-3[2]). Table 1 shows the page hits for two high frequency words, "A" and "The," suggesting that $D \approx 10^{10}$. Table 1 also gives page hits for a couple of low frequency words selected from *The New Oxford Dictionary of English* (Pearsall, 1998), demonstrating that there are lots of hits even for rare words.

How many page hits do "ordinary" words have? To address this question, we randomly picked 15 pages from Learners' dictionary (Hornby, 1989), and selected the first entry on each page. According to Google, there are 10 million pages/word (median value, aggregated over the 15 words). To compute the two-way associations for the 57,100 entries in this dictionary would cost $57100^2 \times 10^7/2 \approx 2^{54}$; three-way associations would cost $57100^3 \times 10^7/6 \approx 2^{68}$; four-way would cost 2^{82}. Clearly, we cannot afford to compute these associations using a straightforward brute force approach.

Estimates are often good enough. We should not have to look at every document to determine whether two words are strongly associated or not. We could use the estimated co-occurrences

∗ Department of Statistics, Stanford, CA 94005
† One Microsoft Way, Redmond, WA 98007
1 This work was conducted at Microsoft while the first author was an intern.
2 All experiments with MSN and Google were conducted in August, 2005.

Table 1
Page hits for a few high frequency and low frequency words.

Query	Hits (MSN)	Hits (Google)
A	2,452,759,266	3,160,000,000
The	2,304,929,841	3,360,000,000
Kalevala	159,937	214,000
Griseofulvin	105,326	149,000
Saccade	38,202	147,000

from a small sample to compute the test statistics, most commonly the Pearson's Chi-squared (χ^2) test, the likelihood ratio (G^2) test, the Fisher's exact test (Dunning, 1993; Agresti, 2002; Manning and Schutze, 1999; Moore, 2004), as well as some non-statistical metrics such as cosine similarity or resemblance (Jaccard coefficient) that are also widely used in computational linguistics and information retrieval (Salton, 1989; Manning and Schutze, 1999; Baeza-Yates and Ribeiro-Neto, 1999).

The conventional sampling method randomly selects D_s documents from a collection of size D and counts the word co-occurrences within the sampled documents. In terms of the term-by-document matrix, which has N rows and D columns, the conventional sampling randomly picks D_s columns. One problem with the conventional sampling is that all words are sampled at the same rate. Word distributions have long tails. There are a few high frequency words and many low frequency words. It would be convenient if the sampling rate could vary from word to word so that the sampling rate would be higher for more interesting words and lower for less interesting words.

Sampling over postings provides a good solution. For each word W, there are a set of postings, P, which contains a set of document IDs, one for each document containing W. In the term-document matrix, each row corresponds to the postings of a specific word. A well-known randomized algorithm that was based on sampling over postings is the "sketch" algorithm developed by Broder (1997), originally motivated to remove nearly-duplicated documents. Broder's sketch algorithm was implemented in Web scale (Broder, 1997; Broder et al., 1997). Broder et al. (1998; Broder et al. (2000) further developed a "minwise" algorithm, which is essentially a "sample-with-replacement" version of the original sketch algorithm.

Charikar (2002) pointed out that Broder's sketch algorithm was a special instance of a *locality sensitive hashing* (LSH) scheme introduced by Indyk and Motwani (1998). Charikar (2002) also re-introduced another LSH scheme,which applied the random projection to estimate cosine similarity, originally proved by Goemans and Williamson (1995). Ravichandran et al. (2005) applied the LSH to generate noun similarity lists from 70 million pages.

Sampling can make it possible to work in memory, avoiding disk. Brin and Page (1998) reported an inverted index of 37.2 GBs for 24 million pages. By extrapolation, we should expect the size of the inverted indexes for current Web scale ($D \approx 10$ billion pages) to be 1500 GBs/billion pages, probably too large for memory. But a sample is more manageable; the inverted index for a 10^{-4} sample of the entire web could fit in memory on a single PC (1.5 GBs).

In estimating the associations, it is desirable that the estimates be consistent. Joint frequencies ought to decrease monotonically as we add terms to the query. Table 2 shows that estimates produced by current search engines are not always consist. Adding a term ("Japan") cannot cause the hits to increase, but the estimates in Table 2 violate this invariant.

Table 2
Estimates of page hits are not always consistent. Joint frequencies ought to decrease monotonically as we add terms to the query, but estimates produced by current state-of-the-art search engines sometimes violate this invariant.

Query	Hits (MSN)	Hits (Google)
America	150,731,182	393,000,000
America, China	15,240,116	66,000,000
America, China, Britain	235,111	6,090,000
America, China, Britain, Japan	154,444	23,300,000

Table 3
For a four word query "Governor, Schwarzenegger, Terminator, Austria," Google returns the estimated document frequencies and all two-way, three-way, and four-way associations. In order to produce the smallest intermediate writes, the optimal order of joins would be: (("Schwarzenegger" $\cap$ "Austria") $\cap$ "Terminator") $\cap$ "Governor," with 136,000 intermediate results. The standard practice starts with the least frequent terms, i.e., (("Schwarzenegger" $\cap$ "Terminator") $\cap$ "Governor") $\cap$ "Austria," with 579,100 intermediate results.

	Query	Hits (Google)
One-way	Austria	88,200,000
	Governor	37,300,000
	Schwarzenegger	4,030,000
	Terminator	3,480,000
Two-way	Governor & Schwarzenegger	1,220,000
	Governor & Austria	708,000
	Schwarzenegger & Terminator	504,000
	Terminator & Austria	171,000
	Governor & Terminator	132,000
	Schwarzenegger & Austria	120,000
Three-way	Governor & Schwarzenegger & Terminator	75,100
	Governor & Schwarzenegger & Austria	46,100
	Schwarzenegger & Terminator & Austria	16,000
	Governor & Terminator & Austria	11,500
Four-way	Governor & Schwarzenegger & Terminator & Austria	6,930

1.1 An Application: The Governator

Table 3 contains estimate of hits for four words and their two-way, three-way, and four-way combinations. Accurate estimates would have applications in Database query planning (Garcia-Molina et al., 2002, Chapter 16). Query optimizers construct a plan to minimize a cost function (e.g., intermediate writes). The optimizer could do better if it could estimate a table like Table 3. But efficiency is important. We certainly don't want to spend more time optimizing the plan than executing it.

Suppose the optimizer wanted to construct a plan for the query: "Governor, Schwarzenegger, Terminator, Austria." The standard solution starts with the least frequent terms: (("Schwarzenegger" $\cap$ "Terminator") $\cap$ "Governor") $\cap$ "Austria." That plan generates 579,100 intermediate writes after the first and second joins. An improvement would be (("Schwarzenegger" $\cap$ "Austria") $\cap$ "Terminator") $\cap$ "Governor," reducing the 579,100 down to 136,000.

In addition to counting hits, Table 3 could also help find the top k pages. When joining the terms, we'd like to know how far down the ranking we should go. Accurate estimates of

associations would help the optimizer make such decisions.

1.2 Sampling and Estimation

	W_2	$\sim W_2$
W_1	a (x_1)	b (x_2)
$\sim W_1$	c (x_3)	d (x_4)

(a) Contingency table

	W_2	$\sim W_2$
W_1	a_s (s_1)	b_s (s_2)
$\sim W_1$	c_s (s_3)	d_s (s_4)

(b) Sample contingency table

Figure 1
(a): A contingency table for word W_1 and word W_2. Cell a is the number of documents that contain both W_1 and W_2, b is the number that contain W_1 but not W_2, c is the number that contain W_2 but not W_1, and d is the number that contain neither W_1 nor W_2. The margins, $f_1 = a + b$ and $f_2 = a + c$ are known as document frequencies in IR. $D = a + b + c + d$ is the total number of documents in the collection. To be consistent with our notation in studying the multi-way associations, a, b, c, and d are also denoted, in parentheses, by x_1, x_2, x_3, and x_4, respectively. (b): A sample contingency table (a_s, b_s, c_s, d_s), where the subscript s indicates the *sample space*. The cells are also numbered as (s_1, s_2, s_3, s_4), to be consistent with multi-way associations.

Two-way associations are often represented as two-way contingency tables (Figure 1(a)). Our task is to construct a sample contingency table (Figure 1(b)), and estimate 1(a) from 1(b). We will use a maximum likelihood estimator (MLE) to find the most likely contingency table, given the sample and various other constraints. We will propose a sampling procedure that bridges two popular choices: (A) sampling over documents and (B) sampling over postings. The estimation task is straightforward and well-understood for (A). As we consider more flexible sampling procedures such as (B), the estimation task becomes more challenging.

We assume a standard inverted index (Witten et al., 1999, section 3.2). For word W_1, there are a set of postings, P_1, containing a set of document IDs, one for each document containing W_1. The size of postings, $f_1 = |P_1|$, corresponds to the margins of the contingency tables in Figure 1(a), also known as document frequencies in IR.

The postings lists are approximated by *sketches*, K, first introduced by Broder (1997) for removing duplicate web pages. Assuming that document IDs are random (e.g., achieved by a random permutation), we can compute K_1, a random sample of P_1, by simply selecting the first few elements of P_1.

In Section 3, we will propose using sketches to construct sample contingency tables. With this novel construction, the contingency table (and summary statistics based on the table) can be estimated using conventional statistical methods such as MLE. This construction can be extended to estimate multi-way associations in a fairly straightforward manner.

2 Broder's Sketch Algorithm

One could randomly sample two postings and intersect the samples to estimate associations. The sketch technique introduced by Broder (1997) is a significant improvement, as demonstrated in Figure 2.

Assume that each document in the corpus of size D is assigned a unique random ID between 1 and D. The postings for word W_1 is a sorted list of f_1 document IDs. The sketch, K_1, is the first (smallest) k_1 document IDs in P_1. Broder used $\text{MIN}_k(Z)$ to denote the k smallest elements in the set, Z. Thus, $K_1 = \text{MIN}_{k_1}(P_1)$. Similarly, P_2 denotes the postings for word W_2, and K_2 denotes its sketch, $\text{MIN}_{k_2}(P_2)$. Broder's algorithm restricted $k_1 = k_2 = k$.

Broder defined resemblance (R) and sample resemblance (R_s) to be:

$$R(\mathrm{W}_1, \mathrm{W}_2) = \frac{\mathrm{P}_1 \cap \mathrm{P}_2}{\mathrm{P}_1 \cup \mathrm{P}_2} = \frac{a}{a+b+c} = \frac{a}{f_1+f_2-a}, \quad (1)$$

$$R_s(\mathrm{W}_1, \mathrm{W}_2) = \frac{|\mathrm{MIN}_k(K_1 \cup \mathrm{K}_2) \cap \mathrm{K}_1 \cap \mathrm{K}_2|}{|\mathrm{MIN}_k(\mathrm{K}_1 \cup \mathrm{K}_2)|}. \quad (2)$$

Broder (1997) proved that R_s is an unbiased estimator of R, i.e., $\hat{R} = R_s$, $\mathrm{E}(\hat{R}) = R$. One could use $\hat{R}$ to estimate a: $\hat{a} = \frac{R_s}{1+R_s}(f_1+f_2)$. However, this is not recommended, for at least three reasons. First, R_s uses only k of the $2 \times k$ samples; smaller samples $\rightarrow$ larger errors. Secondly, the restriction of equal sample size: $k_1 = k_2$, is inflexible and should be removed. Thirdly, the estimate of a from R_s is (slightly) biased because R is not a linear function of a.

In fact, we recommend estimating resemblance from the estimated a using our generalization of the sketch algorithm. Our method does not restrict equal sample size (i.e., $k_1 \neq k_2$ is permitted and actually recommended.), and more effectively uses the samples (can use all $2 \times k$ samples) hence has less errors than Broder's original algorithm.

Before we delve into the details of our algorithm, we present an experiment to show how Broder's sketch improve the coverage of a, as illustrated by Monte Carlo simulations in Figure 2. The figure plots, $\mathrm{E}\left(\frac{a_s}{a}\right)$, percentage of intersections, as a function of (postings) sampling rate, $\frac{k}{f}$, where $f_1 = f_2 = f$, $k_1 = k_2 = k$. The solid lines (sketches), $\mathrm{E}\left(\frac{a_s}{a}\right) \approx \frac{k}{f}$, are above the dashed curve (random sampling), $\mathrm{E}\left(\frac{a_s}{a}\right) = \frac{k^2}{f^2}$. The difference between $\frac{k}{f}$ and $\frac{k^2}{f^2}$ is particularly important at low sampling rates.

To explain the dashed curve in Figure 2, we can analytically show that $\mathrm{E}\left(\frac{a_s}{a}\right) = \frac{k^2}{f^2}$. Suppose $\mathrm{Z} = \mathrm{P}_1 \cap \mathrm{P}_2$, $a = |\mathrm{Z}|$. If we randomly sample k elements from P_1, then by the property of hypergeometric sampling, on average, we have $\frac{k}{f}a$ samples, denoted by Z_1 that belong to the intersection Z, i.e., $\mathrm{E}(|\mathrm{Z}_1|) = \frac{ka}{f}$. Similarly, for k random samples from P_2, we have $\mathrm{E}(|\mathrm{Z}_2|) = \frac{k}{f}a$. By definition, $a_s = |\mathrm{Z}_1 \cap \mathrm{Z}_2|$. Again, by the properties of hypergeometric sampling, we have $\mathrm{E}(a_s) = \frac{1}{a}\left(\frac{k}{f}a\right)\left(\frac{k}{f}a\right) = \frac{k^2}{f^2}a$, i.e., $\mathrm{E}\left(\frac{a_s}{a}\right) = \frac{k^2}{f^2}$.

In contrast, with sketches, we have $\mathrm{E}\left(\frac{a_s}{a}\right) \approx \mathrm{E}\left(\frac{D_s}{D}\right) \approx \frac{k}{f}$. Because the approximate relationship $\mathrm{E}\left(\frac{a_s}{a}\right) \approx \frac{k}{f}$ holds with very good accuracy, we only see one solid curve in the figure.

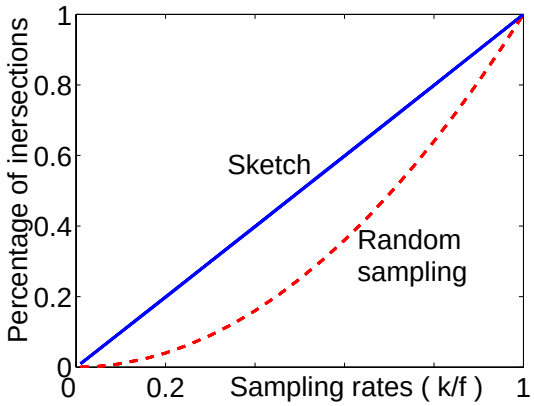


Figure 2
Sketches (solid curves) dominate random sampling (dashed curve). a=0.22, 0.38, 0.65, 0.80, 0.85f, f=0.2D, D=10^5. There is only one dashed curve across all values of a. There are different but indistinguishable solid curves depending on a.

3 Generalizing Sketches, from Resemblance to Contingency Table

Sketches were first proposed for estimating resemblance (R). This section generalizes the method to construct sample contingency tables, from which we can estimate associations and all summary statistics including R and cosine coefficient.

To better explain our construction, we start with an example using conventional random sampling over documents as illustrated in Figure 3. In this example, there are $D = 36$ documents and we choose a corpus sampling rate of 50%, i.e., $D_s = 18$. Since document IDs are assumed to be uniformly random, we can pick any 18 documents to construct a random sample. In particular, we sample the first 18 documents from the collection sorted by document IDs. Suppose we are interested in word W_1 and word W_2, we can construct the sample contingency table from using the $D_s = 18$ samples as in Figure 3.

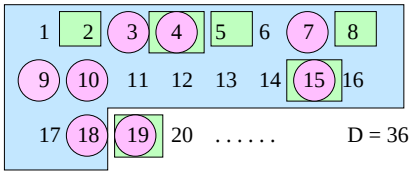


Figure 3
In this corpus, there are $D = 36$ documents numbered from 1 to 36 and sorted ascending. We choose a (corpus) sampling rate of 50%, i.e., $D_s = 18$. Since document IDs are assumed random, we only need to pick the first 18 documents. Suppose we are interested in word W_1 and word W_2. The documents that contain W_1 are marked in small circles, and documents that contain W_2 are in small squares. After we have the samples, we can construct a sample contingency tables for word W_1 and word W_2: a_s = $|\{4, 15\}|$ = 2, b_s = $|\{3, 7, 9, 10, 18\}|$ = 5, c_s = $|\{2, 5, 8\}|$ = 3, d_s = $|\{1, 6, 11, 12, 13, 14, 16, 17\}|$ = 8.

Next, in Figure 4, we present a procedure that uses sketches to construct the same sample contingency table as conventional sampling, using the same example in Figure 3. In this procedure, we sample from the beginning of the postings P_1 and P_2. In order to equivalently sample the first $D_s = 18$ documents, we sample all document IDs in both sketches that are smaller than or equal to 18. After we have the samples, we can then compute a_s, b_s, c_s, and d_s to construct the sample contingency table, which is identical to the example in Figure 3.

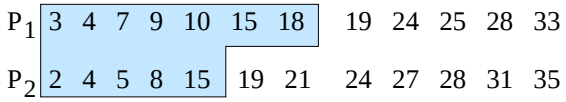


Figure 4
Procedure 1. Suppose we have the same corpus as in Figure 3 and we would like to sample the first $D_s = 18$ documents, to construct a sample contingency table for word W_1 and word W_2, only from their postings P_1 and P_2. As the document IDs in the postings are sorted ascending, we only need to sample from the beginnings of P_1 and P_2 for all documents IDs that are smaller than or equal to $D_s = 18$, as illustrated in the shaded box. In this particular example, this sampling procedure produces a sample contingency table: $a_s = 2$, $b_s = 5$, $c_s = 3$ and $d_s = 8$, identical to the example in Figure 3.

This technique of sampling over postings takes advantage of the fact that the document IDs span the integers from 1 to D with no gaps. When we compare the two sketches that includes all documents IDs smaller than or equal to D_s, we have effectively looked at D_s documents in the original collection.

The above procedure for constructing sketches, however, is not convenient in many situations and is not recommended. When we construct sketches off-line for all words in a corpus, we do not know D_s in advance. In fact, we would like to effectively vary D_s for different word

pairs. For on-line sketch construction (say, only for W_1 and W_2), it is also often much easier to sample according to the postings sampling rate ($\frac{k}{f}$) as opposed to the corpus sampling rate ($\frac{D_s}{D}$), even if we do know the target D_s, because during sampling we certainly do not want to compare samples against D_s.

Next, we will present a slightly different sketch construction procedure that does not require knowing D_s in advance, as illustrated in Figure 5. In this recommended procedure, we build sketches according the postings sampling rates, or equivalently, the pre-specified sketch sizes. When we need to study the associations between two specific words, say W_1 and W_2, we load the sketches K_1 and K_2. The last elements in K_1 and K_2 are respectively denoted as $K_{1(k_1)}$ and $K_{2(k_2)}$, using the standard "order statistics" notation. We treat $D_s = \min\left(K_{1(k_1)}, K_{2(k_2)}\right)$ and trim all documents IDs in K_1 and K_2 that are larger than D_s. Symbolically,

$$
\begin{aligned}
&D_s = \min\{K_{1(k_1)}, K_{2(k_2)}\},\\
&k_1' = k_1 - |\{j : K_{1(j)} > D_s\}|, \quad k_2' = k_2 - |\{j : K_{2(j)} > D_s\}|,\\
&a_s = |K_1 \cap K_2|, \quad b_s = k_1' - a_s, \quad c_s = k_2' - a_s, \quad d_s = D_s - a_s - b_s - c_s.
\end{aligned} \tag{3}
$$

P_1	3	4	7	9	10	15	18	19	24	25	28	33
P_2	2	4	5	8	15	19	21	24	27	28	31	35

Figure 5
Procedure 2. Using the same corpus as in Figures 3 and 4, we illustrate our recommended procedure to construct sample contingency tables from sketches, K_1 and K_2 (larger shaded box). K_1 consists of the first $k_1 = 7$ document IDs in P_1, the postings for word W_1; and K_2 consists of the first $k_2 = 7$ document IDs in P_2, the postings for word W_2. There are 11 document IDs in both W_1 and W_2, and $a = 5$ document IDs in the intersection: {4, 15, 19, 24, 28}. $D_s = \min(18, 21) = 18$. Document IDs 19 and 21 in K_2 are excluded from the sample contingency table because we can not determine if they are in the intersection or not, without looking outside the larger box. As it turns out, 19 is in the intersection and 21 is not. This produce generates a sample contingency table: $a_s = 2$, $b_s = 5$, $c_s = 3$ and $d_s = 8$, which is the same as in Figures 3 and 4.

Although both Procedures 1 (in Figure 4) and Procedure 2 (in Figure 5) produce the same sample contingency tables as the conventional random sampling, they are different in that Procedure 1 requires a pre-specified corpus sample size D_s while Procedure 2 is much more flexible. The analysis for Procedure 1 is the same as for conventional sampling, while the analysis for Procedure 2 is much harder. However, we can see that, *conditional* on D_s, Procedure 2 is the same as Procedure 1. Therefore, to simplify the analysis, our estimation method will be based on conditioning on D_s.

After we have constructed the sample contingency tables, our maximum likelihood estimator (MLE) will estimate the most probable a by solving a cubic MLE equation:

$$
\frac{f_1 - a + 1 - b_s}{f_1 - a + 1} \frac{f_2 - a + 1 - c_s}{f_2 - a + 1} \frac{D - f_1 - f_2 + a}{D - f_1 - f_2 + a - d_s} \frac{a}{a - a_s} = 1. \tag{4}
$$

Assuming "sample-with-replacement," we can have a slightly simpler cubic MLE equation:

$$
\frac{a_s}{a} - \frac{b_s}{f_1 - a} - \frac{c_s}{f_2 - a} + \frac{d_s}{D - f_1 - f_2 + a} = 0, \tag{5}
$$

Instead of solving a cubic equation, we can also estimate a using a very accurate closed-form approximation:

$$
\hat{a} = \frac{f_1\left(2a_s + c_s\right) + f_2\left(2a_s + b_s\right) - \sqrt{\left(f_1\left(2a_s + c_s\right) - f_2\left(2a_s + b_s\right)\right)^2 + 4f_1 f_2 b_s c_s}}{2\left(2a_s + b_s + c_s\right)}. \tag{6}
$$

In Section 5, we will derive the above MLE results in details. We will also analyze the estimation errors, which is directly related to the variance of the estimator. In Section 5, we will derive the following variance formula:

$$\text{Var}\left(\hat{a}\right) \approx \frac{\frac{D}{D_s} - 1}{\frac{1}{a} + \frac{1}{f_1 - a} + \frac{1}{f_2 - a} + \frac{1}{D - f_1 - f_2 + a}}. \tag{7}$$

which is the conditional variance, i.e., in terms of D_s.

In some situations (e.g., for choosing postings sample sizes), we also need the unconditional variance, which is approximated as

$$\text{Var}\left(\hat{a}\right)_{uc} \approx \frac{\max\left(\frac{f_1}{k_1}, \frac{f_2}{k_2}\right) - 1}{\frac{1}{a} + \frac{1}{f_1 - a} + \frac{1}{f_2 - a} + \frac{1}{D - f_1 - f_2 + a}}. \tag{8}$$

Based on statistical large-sample theory, these variance formulas are very accurate when the sketch sizes are reasonable (e.g., $\geq 20 - 50$).

Our sketch construction procedure can be easily extended to estimate multi-way associations, although the MLE solution and variance estimation will be far more complicated. We will present the details for estimating multi-way associations in Section 11. In all other sections, we will focus on two-way associations.

4 Baseline Estimators

Before considering our proposed MLE method, we introduce two baseline estimators that will not work as well. The independence baseline does not take advantage of the samples whereas the margin-free baseline does not take advantage of the margins (e.g., f_1, f_2). It is expected that the margin-free baseline will work better than the (sample-free) independence baseline but our proposed MLE estimator, which takes advantage of both the samples and the margins, will outperform both baselines.

4.1 Independence Baseline

When two words W_1 and W_2 are independent, the size of intersections, a, follows a hypergeometric distribution, i.e.,

$$P(a) = \frac{\binom{f_1}{a}\binom{D-f_1}{f_2-a}}{\binom{D}{f_2}}, \tag{9}$$

with mean and variance (Shao, 1999, Table 1.1)

$$\text{E}(a) = \frac{f_1 f_2}{D}, \qquad \text{Var}(a) = \frac{f_1 f_2 (D - f_1)(D - f_2)}{D^2(D-1)} \leq \frac{f_1 f_2}{D}. \tag{10}$$

Note that (9) is also a common null-hypothesis distribution in testing the independence of a two-way contingency table, i.e., the so-called "Fisher's exact test" (Agresti, 2002, Section 3.5.1).

With the independence assumption, an estimator would be:

$$\hat{a}_{IND} = \text{E}(a) = \frac{f_1 f_2}{D}. \tag{11}$$

Independence assumptions are often made in Databases (Garcia-Molina et al., 2002, Chapter 16.4) and Statistical NLP (Manning and Schutze, 1999, Chapter 13.3).

4.2 Margin-free Baseline

The basic margin-free model is known as the *multivariate hypergeometric* model, which is a generalization of the hypergeometric model and is often illustrated by a simple urn model. Suppose there are D balls in a bag. Each ball has one of the four different colors. There are a, b, c, and d balls broken down by their colors. $D = a + b + c + d$. We pick D_s balls randomly from the bag without replacement and obtain a_s, b_s, c_s, and d_s balls according to the their colors. $D_s = a_s + b_s + c_s + d_s$. The probability of (a_s, b_s, c_s, d_s) follows the multivariate hypergeometric distribution. Note that this model is not limited to 4 cells (colors). This model does not take advantage of the known margins: $f_1 = a + b$, $f_2 = a + c$; hence the name "margin-free model." When the samples a_s, b_s, c_s and d_s are obtained from our sketch construction method, the margin-free model is also based on conditioning on D_s.

The multivariate hypergeometric sample expectations are given by (Siegrist, 1997),

$$\mathrm{E}(a_s) = \frac{D_s}{D}a, \quad \mathrm{E}(b_s) = \frac{D_s}{D}b, \quad \mathrm{E}(c_s) = \frac{D_s}{D}c, \quad \mathrm{E}(d_s) = \frac{D_s}{D}d. \tag{12}$$

And the variance of a_s is

$$\mathrm{Var}(a_s) = D_s\frac{a}{D}\left(1 - \frac{a}{D}\right)\frac{D - D_s}{D - 1}. \tag{13}$$

We can replace a in (13) by b, c, d, to get the variances of b_s, c_s, and d_s, respectively.

Knowing the expectation and variance of the multivariate hypergeometric model allows us to derive an estimator and its variance:

$$\hat{a}_{MF} = \frac{D}{D_s}a_s, \quad \mathrm{Var}(\hat{a}_{MF}) = \frac{D^2}{D_s^2}\mathrm{Var}(a_s) = \frac{D}{D_s}\frac{1}{\frac{1}{a} + \frac{1}{D-a}}\frac{D - D_s}{D - 1}. \tag{14}$$

In the urn model example, if we randomly pick the balls and put them back to the bag after recording their colors, we end up with a *multinomial* model. When the sampling rate $\frac{D_s}{D}$ is low, "sample-with-replacement" is often a good approximation, which in general simplifies the analysis. However, we need to be careful about this assumption since we do not place restrictions on the sampling rate, (allowing for up to 100% sampling of rare words).

For a multinomial distribution, its expectations are the same as given in (12). Its variance, however, is different:

$$\mathrm{Var}(a_s, r) = D_s\frac{a}{D}\left(1 - \frac{a}{D}\right), \tag{15}$$

where the symbol "r" indicates "sample-with-replacement."

According to the multinomial model, an estimator and its variance would be:

$$\hat{a}_{MF,r} = \frac{D}{D_s}a_s, \qquad \mathrm{Var}(\hat{a}_{MF,r}) = \frac{D}{D_s}\frac{1}{\frac{1}{a} + \frac{1}{D-a}}, \tag{16}$$

which implies that, for the margin-free model, the "sample-with-replacement" simplification still results in the same estimator but over-estimates the variance.

In (14), the term $\frac{D-D_s}{D-1} \approx \frac{D-D_s}{D}$, which varies from unity (at 0% sampling rate) to zero (at 100% sample rate), is often called the "finite population correction factor" (Siegrist, 1997).

5 The Proposed MLE Method

The task is to estimate the contingency table from the samples, the margins and D. We would like to use a maximum likelihood estimator for the most probable a, which maximizes the

(full) likelihood (probability mass function, PMF) $P(a_s, b_s, c_s, d_s; a)$. Unfortunately, we do not know the exact expression for $P(a_s, b_s, c_s, d_s; a)$, but we do know the conditional probability $P(a_s, b_s, c_s, d_s|D_s; a)$. Since the document IDs are uniformly random, sampling the first D_s contiguous documents is statistically equivalent to randomly sampling D_s documents from the corpus. Based on this key observation and Figure 5, *conditional* on D_s, $P(a_s, b_s, c_s, d_s|D_s; a)$ is the PMF of a two-way sample contingency table.

We factor the full likelihood into:

$$P(a_s, b_s, c_s, d_s; a) = P(a_s, b_s, c_s, d_s|D_s; a) \times P(D_s; a), \tag{17}$$

where $P(a_s, b_s, c_s, d_s|D_s; a)$ is the likelihood of the conditional sample contingency table. The marginal probability $P(D_s; a)$ is difficult. However, since we do not expect a strong dependency of D_s on a (as illustrated in Figure 6), we use a *partial likelihood*, which seeks the a that maximizes the partial likelihood $P(a_s, b_s, c_s, d_s|D_s; a)$ instead of the full probability. The partial likelihood method is widely used in statistics. A well-known example would be the Cox proportional hazards model in survival analysis (Venables and Ripley, 2002, Section 13.3).

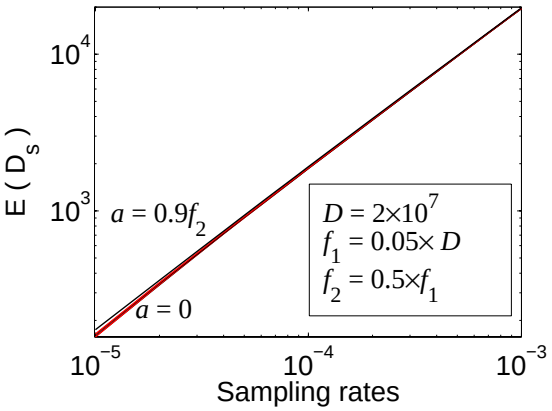


Figure 6
This experiment shows that $\mathrm{E}(D_s)$ is not sensitive to a. $D = 2 \times 10^7$, $f_1 = D/20$, $f_2 = f_1/2$. The different curves correspond to a = 0, 0.05, 0.2, 0.5 and 0.9 f_2. These curves are almost indistinguishable except at very low sampling rates. Note that, at sampling rate = 10^{-5}, the sample size $k_2 = 5$ only.

Conditional on D_s, the partial likelihood is

$$\begin{aligned} P(a_s, b_s, c_s, d_s|D_s; a) &= \frac{\binom{a}{a_s}\binom{b}{b_s}\binom{c}{c_s}\binom{d}{d_s}}{\binom{a+b+c+d}{a_s+b_s+c_s+d_s}} = \frac{\binom{a}{a_s}\binom{f_1-a}{b_s}\binom{f_2-a}{c_s}\binom{D-f_1-f_2+a}{d_s}}{\binom{D}{D_s}} \\ &\propto \frac{a!}{(a-a_s)!} \times \frac{(f_1-a)!}{(f_1-a-b_s)!} \times \frac{(f_2-a)!}{(f_2-a-c_s)!} \times \frac{(D-f_1-f_2+a)!}{(D-f_1-f_2+a-d_s)!} \\ &= \prod_{i=0}^{a_s-1}(a-i) \times \prod_{i=0}^{b_s-1}(f_1-a-i) \times \prod_{i=0}^{c_s-1}(f_2-a-i) \times \prod_{i=0}^{d_s-1}(D-f_1-f_2+a-i), \end{aligned} \tag{18}$$

where $\binom{n}{m} = \frac{n!}{m!(n-m)!}$. "$\propto$" denotes "proportional to." The multiplicative terms not mentioning a are discarded, because they can be considered as constants and will not contribute to the MLE. To the best of our knowledge, there are no known MLE results for (18).

Let $\hat{a}_{MLE}$ be the value of a that maximizes the partial likelihood (18), or equivalently,

maximizes the $\log$[1] likelihood, $\log P\left(a_s, b_s, c_s, d_s | D_s; a\right)$:

$$\sum_{i=0}^{a_s-1} \log(a-i) + \sum_{i=0}^{b_s-1} \log\left(f_1-a-i\right) + \sum_{i=0}^{c_s-1} \log\left(f_2-a-i\right) + \sum_{i=0}^{d_s-1} \log\left(D-f_1-f_2+a-i\right),$$

whose first derivative, $\frac{\partial \log P(a_s,b_s,c_s,d_s|D_s;a)}{\partial a}$, is

$$\sum_{i=0}^{a_s-1} \frac{1}{a-i} - \sum_{i=0}^{b_s-1} \frac{1}{f_1-a-i} - \sum_{i=0}^{c_s-1} \frac{1}{f_2-a-i} + \sum_{i=0}^{d_s-1} \frac{1}{D-f_1-f_2+a-i}. \tag{19}$$

Since the second derivative, $\frac{\partial^2 \log P(a_s,b_s,c_s,d_s|D_s;a)}{\partial a^2}$,

$$-\sum_{i=0}^{a_s-1} \frac{1}{(a-i)^2} - \sum_{i=0}^{b_s-1} \frac{1}{(f_1-a-i)^2} - \sum_{i=0}^{c_s-1} \frac{1}{(f_2-a-i)^2} - \sum_{i=0}^{d_s-1} \frac{1}{(D-f_1-f_2+a-i)^2},$$

is negative, we know that the log likelihood function is concave, and therefore, there is a unique maximum. One could use some numerical methods to solve (19) for $\frac{\partial \log P(a_s,b_s,c_s,d_s|D_s;a)}{\partial a} = 0$, which is quite complex and may subject to numerical difficulties.

It turns out that we can derive an exact (and much simpler) solution by developing the following updating formula from (18):

$$\begin{aligned}
&P\left(a_s, b_s, c_s, d_s | D_s; a\right) \\
=&P\left(a_s, b_s, c_s, d_s | D_s; a-1\right) \times \frac{a}{a-a_s} \frac{f_1-a+1-b_s}{f_1-a+1} \frac{f_2-a+1-c_s}{f_2-a+1} \frac{D-f_1-f_2+a}{D-f_1-f_2+a-d_s} \\
=&P\left(a_s, b_s, c_s, d_s | D_s; a-1\right) \times g(a).
\end{aligned} \tag{20}$$

Since we know that the MLE exists and is unique, it suffices to find the a from $g(a) = 1$,

$$g(a) = \frac{a}{a-a_s} \frac{f_1-a+1-b_s}{f_1-a+1} \frac{f_2-a+1-c_s}{f_2-a+1} \frac{D-f_1-f_2+a}{D-f_1-f_2+a-d_s} = 1, \tag{21}$$

which is cubic in a (the fourth term vanishes), solved either exactly or numerically. The well-known Cardano formula can be used to solve this cubic equation (Weisstein, 2005b, Web resource). However, we recommend a numerical method, which appears to be much simpler and more straightforward.

$g(a) = 1$ is equivalent to $q(a) = \log g(a) = 0$. The first derivative of $q(a)$ is

$$\begin{aligned}
q'(a) = &\left(\frac{1}{f_1-a+1} - \frac{1}{f_1-a+1-b_s}\right) + \left(\frac{1}{f_2-a+1} - \frac{1}{f_2-a+1-c_s}\right) \\
&+ \left(\frac{1}{D-f_1-f_2+a} - \frac{1}{D-f_1-f_2+a-d_s}\right) + \left(\frac{1}{a} - \frac{1}{a-a_s}\right).
\end{aligned} \tag{22}$$

We can solve for $q(a) = 0$ iteratively using the Newton's method,

$$a^{(\text{new})} = a^{(\text{old})} - \frac{q(a^{(\text{old})})}{q'(a^{(\text{old})})}. \tag{23}$$

In Appendix A, we provide the C code that implements the Newoton's method as described.

1 $\log$ always denotes logarithm with base e, i.e., the natural $\log$.

5.1 The "Sample-with-replacement" Simplification

Under the "sample-with-replacement" assumption, the likelihood function is slightly simpler:

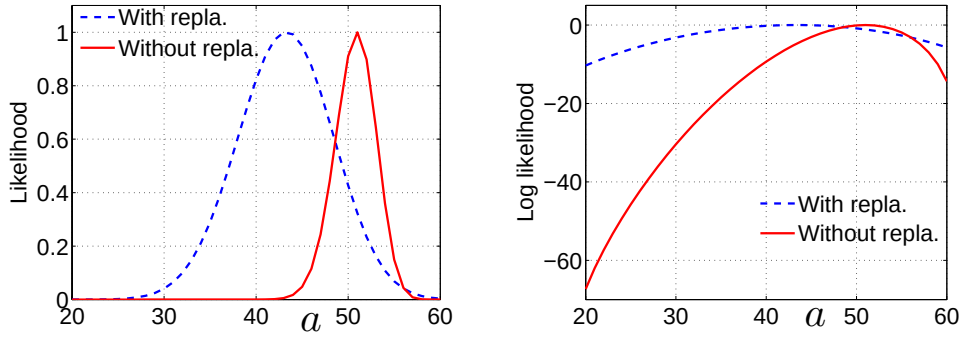
$$P(a_s, b_s, c_s, d_s | D_s; a, r) = \binom{D_s}{a_s, b_s, c_s, d_s} \left(\frac{a}{D}\right)^{a_s} \left(\frac{b}{D}\right)^{b_s} \left(\frac{c}{D}\right)^{c_s} \left(\frac{d}{D}\right)^{d_s}$$
$$\propto a^{a_s} (f_1 - a)^{b_s} (f_2 - a)^{c_s} (D - f_1 - f_2 + a)^{d_s}. \quad (24)$$

Setting the first derivative of the log likelihood to be zero, we can get the following equation

$$\frac{a_s}{a} - \frac{b_s}{f_1 - a} - \frac{c_s}{f_2 - a} + \frac{d_s}{D - f_1 - f_2 + a} = 0, \quad (25)$$

which is also a cubic equation and is slightly simpler than solving $g(a) = 1$.

As shown in Section 4.2, using the margin-free model, the "sample-with-replacement" assumption amplifies the variance but does not change the estimations. With our proposed MLE, the "sample-with-replacement" assumption will change the estimations, although in general we do not expect big differences. Figure 7 gives an example. The figure shows the concavity of the log likelihood and indicates that assuming "sample-with-replacement" may result in a wider log likelihood profile, i.e., wider confidence interval and numerically harder to locate the peak (solution of MLE).



(a) (b)

Figure 7
An example: a_s = 20, b_s = 40, c_s = 40, d_s = 800, $f_1 = f_2$ = 100, D = 1000. The estimated $\hat{a}$ = 43 for "sample-with-replacement," and $\hat{a}$ = 51 for "sample-without-replacement." (a): The likelihood profile, normalized to have a maximum = 1. (b): The log likelihood profile, normalized to have a maximum = 0.

5.2 A Convenient Practical Approximation

Solving a cubic equation for the exact MLE may be so inconvenient that some people may prefer the less accurate margin-free baseline because of its simplicity. This section derives a convenient approximation to the exact MLE.

The idea is to assume that one can identify a_s from K_1 without the knowledge of K_2. In other words, we assume that a_s is hypergeometrically distributed in K_1. Similarly, a_s is also hypergeometrically distributed in K_2, and is independent of the sample in K_1.

Further assuming "sample-with-replacement," a_s is then binomially distributed, $a_s \sim \mathrm{Binom}(a_s + b_s, \frac{a}{f_1})$. Similarly, assume $a_s \sim \mathrm{Binom}(a_s + c_s, \frac{a}{f_2})$. Under these assumptions, the PMF of a_s is

a product of two binomial PMFs:

$$\left[\binom{f_1}{a_s+b_s}\left(\frac{a}{f_1}\right)^{a_s}\left(\frac{f_1-a}{f_1}\right)^{b_s}\right]\times\left[\binom{f_2}{a_s+c_s}\left(\frac{a}{f_2}\right)^{a_s}\left(\frac{f_2-a}{f_2}\right)^{c_s}\right]$$
$$\propto a^{2a_s}\left(f_1-a\right)^{b_s}\left(f_2-a\right)^{c_s}. \tag{26}$$

Setting the first derivative of the logarithm of (26) to be zero, we obtain

$$\frac{2a_s}{a}-\frac{b_s}{f_1-a}-\frac{c_s}{f_2-a}=0, \tag{27}$$

which is quadratic in a, i.e.,

$$a^2(2a_s+b_s+c_s)-a\left(f_1(2a_s+c_s)+f_2(2a_s+b_s)\right)+2a_sf_1f_2=0, \tag{28}$$

and has a convenient closed-form solution:

$$\hat{a}_{MLE,a}=\frac{f_1\left(2a_s+c_s\right)+f_2\left(2a_s+b_s\right)}{2\left(2a_s+b_s+c_s\right)}$$
$$-\frac{\sqrt{\left(f_1\left(2a_s+c_s\right)-f_2\left(2a_s+b_s\right)\right)^2+4f_1f_2b_sc_s}}{2\left(2a_s+b_s+c_s\right)}. \tag{29}$$

The second root can be ignored because it is always out of range:

$$\frac{f_1\left(2a_s+c_s\right)+f_2\left(2a_s+b_s\right)+\sqrt{\left(f_1\left(2a_s+c_s\right)-f_2\left(2a_s+b_s\right)\right)^2+4f_1f_2b_sc_s}}{2\left(2a_s+b_s+c_s\right)}$$
$$\geq\frac{f_1\left(2a_s+c_s\right)+f_2\left(2a_s+b_s\right)+\left|f_1\left(2a_s+c_s\right)-f_2\left(2a_s+b_s\right)\right|}{2\left(2a_s+b_s+c_s\right)}$$
$$\geq\begin{cases}f_1 & \text{if } f_1\left(2a_s+c_s\right)\geq f_2\left(2a_s+b_s\right)\\ f_2 & \text{if } f_1\left(2a_s+c_s\right)<f_2\left(2a_s+b_s\right)\end{cases}$$
$$\geq\min(f_1,f_2).$$

Our evaluations in Section 6 will show that $\hat{a}_{MLE,a}$ is very close to $\hat{a}_{MLE}$.

Now we can examine why the "sample-with-replacement" assumption is necessary in order to obtain a quadratic equation. Without assuming "sample-with-replacement," the approximate PMF would be:

$$P(a_s,b_s,c_s,d_s;D_s,a)_{approx}=\left[\frac{\binom{a}{a_s}\binom{f_1-a}{b_s}}{\binom{f_1}{a_s+b_s}}\right]\left[\frac{\binom{a}{a_s}\binom{f_2-a}{b_s}}{\binom{f_2}{b_s+c_s}}\right]$$
$$\propto\left(\prod_{j=0}^{a_s-1}(a-j)\right)^2\prod_{j=0}^{b_s-1}(f_1-a-j)\prod_{j=0}^{c_s-1}(f_2-a-j). \tag{30}$$

An updating formula would be:

$$P(a_s,b_s,c_s,d_s;D_s,a)_{approx}=P(a_s,b_s,c_s,d_s;D_s,a-1)_{approx}\times$$
$$\left(\frac{a}{a-a_s}\right)^2\frac{f_1-a+1-b_s}{f_1-a+1}\frac{f_2-a+1-c_s}{f_2-a+1}. \tag{31}$$

Therefore, it suffices to find the a such that

$$\left(\frac{a}{a-a_s}\right)^2\frac{f_1-a+1-b_s}{f_1-a+1}\frac{f_2-a+1-c_s}{f_2-a+1}=1, \tag{32}$$

which is still a cubic equation.

5.3 Theoretical Evaluation: Conditional Variance and Bias

How good are the estimates? A popular metric is the mean square error (MSE):

$$\text{MSE}(\hat{a}) = \text{E}\left(\hat{a} - a\right)^2 = \text{E}(\hat{a} - \text{E}(\hat{a}))^2 + (\text{E}(\hat{a}) - a)^2 = \text{Var}\left(\hat{a}\right) + \text{Bias}^2\left(\hat{a}\right). \quad (33)$$

If $\hat{a}$ is unbiased, i.e., $\text{E}(\hat{a}) = a$, then $\text{MSE}(\hat{a}) = \text{Var}\left(\hat{a}\right) = \text{SE}^2\left(\hat{a}\right)$, where $\text{SE}\left(\hat{a}\right)$ is the standard error. Here, all expectations are conditioned on D_s. In general, MLE is biased, i.e., $\text{E}(\hat{a}) \neq a$. However, the large sample theory (Lehmann and Casella, 1998, Theorem 6.3.10) says that, assuming "sample-with-replacement," $\hat{a}_{MLE}$ is asymptotically unbiased and converges to a Normal with mean a and variance $\frac{1}{\text{I}(a)}$, i.e.,

$$\hat{a}_{MLE} \stackrel{D}{\Longrightarrow} \text{N}\left(a, \frac{1}{\text{I}(a)}\right), \quad (34)$$

where $\text{I}(a)$, the (Expected) Fisher Information, is

$$\text{I}(a) = -\text{E}\left(\frac{\partial^2}{\partial a^2} \log P\left(a_s, b_s, c_s, d_s; a, r\right)\right). \quad (35)$$

There are a few issues we should keep in mind when applying the large sample theory.

1. The large sample theory assumes that the samples are i.i.d. In other words, we have to assume "sample-with-replacement" in order to apply the large sample theory.

2. To apply the asymptotic results, the sample size has to be large "enough," but there are no clear cut-off how large is large enough. Since we are working with large corpora, the sample size is in general not an issue. In our experiments, when the sample size is $\geq 20 - 50$, the large sample theory can give quite accurate results.

3. With very small samples, $\frac{1}{\text{I}(a)}$ will under-estimate the variance, by the Information Inequality:$\text{Var}(\hat{a}_{MLE}) \geq \frac{1}{\text{I}(a)}$ (Lehmann and Casella, 1998, Theorem 2.5.10).

4. The asymptotic distribution is represented by a Normal, whose support ranges the whole real line, i.e., $(-\infty, \infty)$, although we know that a can not be negative, nor can it be larger than $\min(f_1, f_2)$. This may not be a concern in most cases, but sometimes we do not need to take it into consideration. In fact, the asymptotic distribution can be represented by other asymptotically equivalent distributions such as Gamma or Beta, as long as their first two moments are matched (either identical, or only asymptotically equivalent) (Li et al., 2005).

5. The expectation in (35) is usually difficult to evaluate. One can compute it numerically (e.g., by Monte Carlo simulations), or resort to some reasonable approximations. Alternatively, one can use the "Observed Fisher Information," i.e., without evaluating the expectation in (35). In fact, the Observed Fisher Information is often considered more reasonable as a measure of variability (Efron and Hinkley, 1978; Siegmund, 1985, Chapter III.9) for many applications such as sequential analysis. However, for performance evaluations or for choosing sample sizes, we will still need to compute the Expected Fisher Information.

Because we would still like to consider the beneficial aspect that our algorithm is "sample-without-replacement," we will correct the asymptotic variance $\frac{1}{\text{I}(a)}$ by multiplying it with the finite population correction factor $\frac{D - D_s}{D}$.

Assuming "sample-with-replacement," the second derivative of the PMF is

$$\frac{\partial^2 \log P(a_s, b_s, c_s, d_s | D_s, a; r)}{\partial a^2} = -\left(\frac{a_s}{a^2} + \frac{b_s}{(f_1 - a)^2} + \frac{c_s}{(f_2 - a)^2} + \frac{d_s}{(D - f_1 - f_2 + a)^2}\right) \tag{36}$$

The Observed Fisher Information would be:

$$\mathrm{I}(a)_{obs} = \frac{a_s}{a^2} + \frac{b_s}{(f_1 - a)^2} + \frac{c_s}{(f_2 - a)^2} + \frac{d_s}{(D - f_1 - f_2 + a)^2}, \tag{37}$$

and the Expected Fisher Information would be:

$$\begin{aligned}\mathrm{I}(a) &= \frac{\mathrm{E}(a_s)}{a^2} + \frac{\mathrm{E}(b_s)}{(f_1 - a)^2} + \frac{\mathrm{E}(c_s)}{(f_2 - a)^2} + \frac{\mathrm{E}(d_s)}{(D - f_1 - f_2 + a)^2} \\ &\approx \frac{D_s}{D}\left(\frac{1}{a} + \frac{1}{f_1 - a} + \frac{1}{f_2 - a} + \frac{1}{D - f_1 - f_2 + a}\right).\end{aligned} \tag{38}$$

Since no closed-form $\mathrm{E}(a_s)$, $\mathrm{E}(b_s)$, $\mathrm{E}(c_s)$, $\mathrm{E}(d_s)$ exist, we plug (12) from the margin-free model into (38) as an approximation. We consider the errors due to this approximation to be "second-order."

To this end, we have obtained an approximate variance of $\hat{a}_{MLE}$,

$$\mathrm{Var}\left(\hat{a}_{MLE}\right) \approx \frac{1}{\mathrm{I}(a)}\frac{D - D_s}{D} = \frac{\frac{D}{D_s} - 1}{\frac{1}{a} + \frac{1}{f_1 - a} + \frac{1}{f_2 - a} + \frac{1}{D - f_1 - f_2 + a}}. \tag{39}$$

Comparing (16) with (39), we can see that $\mathrm{Var}\left(\hat{a}_{MLE}\right) < \mathrm{Var}\left(\hat{a}_{MF}\right)$, as expected. In addition, $\hat{a}_{MLE}$ is (conditionally) asymptotically unbiased while $\hat{a}_{MF}$ is no longer unbiased under margin constraints. Therefore, we expect $\hat{a}_{MLE}$ has smaller MSE than $\hat{a}_{MF}$.

We call the variance computed using the Observed Fisher Information as the "observed variance":

$$\mathrm{Var}\left(\hat{a}\right)_{obs} = \frac{1 - \frac{D_s}{D}}{\frac{a_s}{a^2} + \frac{b_s}{(f_1 - a)^2} + \frac{c_s}{(f_2 - a)^2} + \frac{d_s}{(D - f_1 - f_2 + a)^2}}. \tag{40}$$

5.4 Unconditional Bias and Variance

$\hat{a}_{MLE}$ is also (practically) unconditionally unbiased:

$$\mathrm{E}\left(\hat{a}_{MLE} - a\right) = \mathrm{E}\left(\mathrm{E}\left(\hat{a}_{MLE} - a | D_s\right)\right) \approx \mathrm{E}(0) = 0. \tag{41}$$

The unconditional variance is useful because often we would like to estimate the errors before knowing D_s (e.g., for choosing sample sizes). The unconditional variance can be computed using the following well-known conditional variance formula (Ross, 2002, Chapter 7.4.4):

$$\begin{aligned}\mathrm{Var}\left(\hat{a}_{MLE}\right)_{uc} &= \mathrm{E}\left(\mathrm{Var}\left(\hat{a}_{MLE}\left|\frac{D}{D_s}\right.\right)\right) + \mathrm{Var}\left(\mathrm{E}\left(\hat{a}_{MLE}\left|\frac{D}{D_s}\right.\right)\right) \\ &\approx \frac{\mathrm{E}\left(\frac{D}{D_s}\right) - 1}{\frac{1}{a} + \frac{1}{f_1 - a} + \frac{1}{f_2 - a} + \frac{1}{D - f_1 - f_2 + a}},\end{aligned} \tag{42}$$

because of the conditional asymptotic unbiasedness of $\hat{a}$: $\mathrm{E}\left(\hat{a}_{MLE}|\frac{D}{D_s}\right) \approx a$, which is a constant, hence $\mathrm{Var}\left(\mathrm{E}\left(\hat{a}_{MLE}|\frac{D}{D_s}\right)\right) \approx 0$. Note that for estimating the unconditional variance of

$\hat{a}_{MF}$ by replacing $\frac{D}{D_s}$ with $\mathrm{E}\left(\frac{D}{D_s}\right)$ will under-estimate the variance because of the conditional bias in $\hat{a}_{MF}$.

To evaluate $\mathrm{E}\left(\frac{D}{D_s}\right)$ exactly, we need the exact PMF $P(D_s)$, which we do not know. Even if $P(D_s)$ were available, we would still need to evaluate $\mathrm{E}\left(\frac{D}{D_s}\right)$ numerically, especially because $\frac{D}{D_s}$ is a reciprocal function of D_s. We will resort to approximate solutions.

Recall $D_s = \min\left(\mathrm{K}_{1(k_1)}, \mathrm{K}_{2(k_2)}\right)$. $\mathrm{K}_{1(k_1)}$ is the order statistics of discrete random variables in $[1, D]$ (Siegrist, 1997) with PMF and expectations

$$P(\mathrm{K}_{1(k_1)} = t) = \frac{\binom{t-1}{k_1-1}\binom{D-t}{f_1-k_1}}{\binom{D}{f_1}}, \tag{43}$$

$$\mathrm{E}\left(\mathrm{K}_{1(k_1)}\right) = \frac{k_1(D+1)}{f_1+1} \approx \frac{k_1}{f_1}D, \tag{44}$$

$$\mathrm{Var}\left(\mathrm{K}_{1(k_1)}\right) = \frac{(D+1)(D-f_1)k_1(f_1+1-k_1)}{(f_1+1)^2(f_1+2)}. \tag{45}$$

Alternatively, if we approximate $\frac{\mathrm{K}_{1(k_1)}}{D}$ as a continuous random variable on [0,1], then the density function of $\frac{\mathrm{K}_{1(k_1)}}{D}$ would be (Ross, 2002, Section 6.6)

$$P\left(\frac{\mathrm{K}_{1(k_1)}}{D} = t\right) = \frac{\Gamma(f_1+1)}{\Gamma(f_1-k_1+1)\Gamma(k_1)}t^{k_1-1}(1-t)^{f_1-k_1}, \tag{46}$$

which is a Beta distribution with parameters (k_1, f_1-k_1+1), with mean and variance (Shao, 1999, Table 1.2)

$$\mathrm{E}\left(\frac{\mathrm{K}_{1(k_1)}}{D}\right) = \frac{k_1}{k_1+f_1-k_1+1} = \frac{k_1}{f_1+1} \approx \frac{k_1}{f_1}, \tag{47}$$

$$\mathrm{Var}\left(\frac{\mathrm{K}_{1(k_1)}}{D}\right) = \frac{k_1(f_1-k_1+1)}{(f_1+2)(f_1+1)^2}. \tag{48}$$

We can see that both the exact (discrete) PMF or the continuous approximation imply that

$$\mathrm{E}\left(\mathrm{K}_{1(k_1)}\right) = \frac{k_1}{f_1}D, \quad \mathrm{E}\left(\mathrm{K}_{2(k_2)}\right) = \frac{k_2}{f_2}D. \tag{49}$$

The $\min$ function can be considered concave. By Jensen's inequality (see Cover and Thomas (1991, Theorem 2.6.2) or Durrett (1995, Section 1.3.a)), we know that

$$\begin{aligned}\mathrm{E}\left(\frac{D_s}{D}\right) &= \mathrm{E}\left(\min\left(\frac{\mathrm{K}_{1(k_1)}}{D}, \frac{\mathrm{K}_{2(k_2)}}{D}\right)\right) \\ &\leq \min\left(\frac{\mathrm{E}(\mathrm{K}_{1(k_1)})}{D}, \frac{\mathrm{E}(\mathrm{K}_{2(k_2)})}{D}\right) = \min\left(\frac{k_1}{f_1}, \frac{k_2}{f_2}\right).\end{aligned} \tag{50}$$

The reciprocal function, is convex. Again by Jensen's inequality, we have

$$\mathrm{E}\left(\frac{D}{D_s}\right) = \mathrm{E}\left(\frac{1}{D_s/D}\right) \geq \frac{1}{\mathrm{E}\left(\frac{D_s}{D}\right)} \geq \max\left(\frac{f_1}{k_1}, \frac{f_2}{k_2}\right). \tag{51}$$

By replacing the inequalities with equalities, we use the following approximations:

$$\mathrm{E}\left(\frac{D_s}{D}\right) \approx \min\left(\frac{k_1}{f_1}, \frac{k_2}{f_2}\right), \tag{52}$$

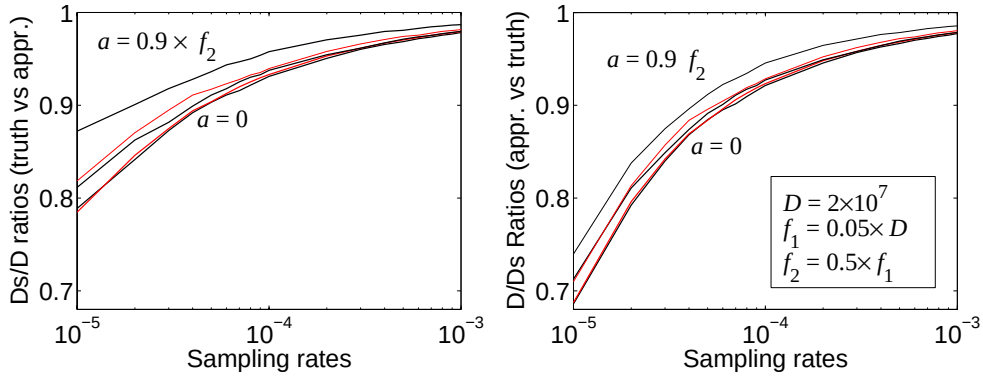
$$\mathrm{E}\left(\frac{D}{D_s}\right) \approx \max\left(\frac{f_1}{k_1}, \frac{f_2}{k_2}\right). \tag{53}$$

Equation (52) gives a very intuitive relationship between the corpus sampling rate $\frac{D_s}{D}$ and the postings sampling rate $\frac{k_1}{f_1}$ ($\frac{k_2}{f_2}$).

With (53), our approximate unconditional variance formula would be:

$$\text{Var}\left(\hat{a}_{MLE}\right)_{uc} \approx \frac{\max\left(\frac{f_1}{k_1}, \frac{f_2}{k_2}\right) - 1}{\frac{1}{a} + \frac{1}{f_1 - a} + \frac{1}{f_2 - a} + \frac{1}{D - f_1 - f_2 + a}}. \tag{54}$$

In our experiment, the approximation in (52) and (53) work well. For example, when the sample size is ≥ 50, the errors in (52) and (53) are usually $\leq 5\% - 10\%$. See the example in Figure 8.



(a) $\frac{D_s}{D} / \min\left(\frac{k_1}{f_1}, \frac{k_2}{f_2}\right)$ (b) $\max\left(\frac{f_1}{k_1}, \frac{f_2}{k_2}\right) / \frac{D}{D_s}$

Figure 8

This is the same experiment in Figure 6. $D = 2 \times 10^7$, $f_1 = 0.05D$, $f_2 = f_1/2$. The different curves are for a = 0, 0.05, 0.2, 0.5 and $0.9 f_2$. (a) plots $\frac{D_s}{D} / \min\left(\frac{k_1}{f_1}, \frac{k_2}{f_2}\right)$ and (b) plots $\max\left(\frac{f_1}{k_1}, \frac{f_2}{k_2}\right) / \frac{D}{D_s}$. We can see that at sampling rates $\geq 10^{-4}$, i.e., $k_2 \geq 50$, the approximations in (52) and (53) work well.

We consider (52) and (53) to be simple, intuitive, and accurate enough in practice. As shown in the experiment in Figure 6, $\text{E}(D_s)$ is not sensitive to a. One approach to improve the accuracy in estimating $\text{E}\left(\frac{D_s}{D}\right)$ and $\text{E}\left(\frac{D}{D_s}\right)$ is to assume independence between W_1 and W_2. The independence assumption will result in exact (but also very sophisticated) solutions for $\text{E}\left(\frac{D_s}{D}\right)$ and $\text{E}\left(\frac{D}{D_s}\right)$, which have to be evaluated numerically.

5.5 Smoothing

The classical smoothing (also frequently referred to as "discounting") methods assume some "prior" distribution on the cells (a, b, c, and d). A commonly-used prior distribution for multinomial sampling is the well-known Dirichlet prior, which is the conjugate prior for multinomial distributions (Gelman et al., 2004, Chapter 3.5). See Teh et al. (2004) for some NLP applications of the Dirichlet prior and Dirichlet process.

For the convenience of extending the discussions to multi-way associations, we use the alternative notation in Figure 1, i.e., (x_1, x_2, x_3, x_4) for (a,b,c,d), and (s_1, s_2, s_3, s_4) for (a_s,b_s,c_s,d_s). In addition, we use N for the total number of cells. $N = 4$ for two-way associations and $N = 2^m$ for multi-way associations where m is the number of words.

The Dirichlet prior can be written as

$$P(\pi_i, 1 \leq i \leq N) = \frac{\Gamma\left(\sum_{i=1}^N \gamma_i\right)}{\prod_{i=1}^N \Gamma\left(\gamma_i\right)} \prod_{i=1}^N \pi_i^{\gamma_i - 1}, \tag{55}$$

where $\pi_i = \frac{x_i}{D}$, $\Gamma(.)$ is the Gamma function .

A particularly popular Dirichlet prior is to let $\gamma_1 = \gamma_2 = ... = \gamma_N = 1$. With this choice, the Bayes estimator for x_i (assuming multinomial) would be:

$$\hat{x}_{i,MF+S} = \frac{s_i + 1}{D_s + N} D, \tag{56}$$

where "S" standards for "smoothing." With this particular prior, the Bayes estimator is effectively adding one to each cell of observations, hence the name "Add-one smoothing" (sometimes also called "Laplace pseudo-count" (Manning and Schutze, 1999, Section 6.2.2)).

We will approximately use the "add-one" law for smoothing the exact MLE estimations. In other words, we will add one to each cell of the observations and plug in the modified observations for the smoothed MLE solutions.

The "add-one" rule has been criticized for the poor performance in estimating n-gram language models (Church and Gale, 1991)(Manning and Schutze, 1999, Section 6.2.5). The problem is that "one" is too big in many applications. In a bigram model, for an example, the probability that two words are adjacent is very small for most words. In (Church and Gale, 1991), when the add-one rule is used, 46.5% of the probability space has actually been assigned to unseen bigrams, which is way too much. The good current practice in NLP in estimating the n-gram model is to use the Good-Turing smoothing and linear interpolation or back-off (Good, 1953; Church and Gale, 1991; Katz, 1987)(Manning and Schutze, 1999, Section 6.2-6.4). Extensive evaluations of a variety of smoothing methods can be found in (Chen and Goodman, 1996; Chen and Goodman, 1998).

The word association is somewhat different from the n-gram model case. Instead of estimating the strict (ordered) adjacency probabilities, we are estimating the the co-occurrences, which normally have much larger probability, i.e., the impact of "add-one" should be much smaller. One convenient property of the "add-one" rule is that one can smooth the estimations on a per-pair base. In this study, for simplicity, we will report the performance of the "add-one" smoothing. In fact, our evaluations show that "add-one" smoothing does not improve the margin-free estimator, especially for two-way associations. However, the "add-one" rule works surprising well for improving the exact MLE, which considers the margins.

We can give some theoretical analysis on the impact, in terms of variance and bias, of the "add-one" smoothing on the performance of the margin-free estimator.

The variance of $\hat{x}_{i,MF+S}$ would be

$$\text{Var}(\hat{x}_{i,MF+S}) = \text{Var}\left(\frac{D(s_i+1)}{D_s+N}\right) = \text{Var}\left(\frac{Ds_i}{D_s+N}\right) \approx \text{Var}\left(\frac{Ds_i}{D_s}\right) = \text{Var}(\hat{x}_{i,MF}), \tag{57}$$

unless N is large. The squared bias would be

$$\begin{aligned}\text{Bias}^2(\hat{x}_{i,MF+S}) &= \left(\text{E}\left(\frac{D(s_i+1)}{D_s+N}\right) - x_i\right)^2 = \left(\frac{D\text{E}(s_i)+D}{D_s+N} - x_i\right)^2 \\ &\approx \left(\frac{D_s x_i + D}{D_s+N} - x_i\right)^2 = \left(\frac{D - x_i N}{D_s+N}\right)^2,\end{aligned} \tag{58}$$

which could be substantial, unless x_i is large or N is large.

In two-way associations, $N = 4$ is small. Therefore, it is possible that "add-one" smoothing will only increase bias without decreasing the variance. Our experiments will verify that $\hat{a}_{MF+S}$ hurts the performance of $\hat{a}_{MF}$ in most cases. In multi-way associations, since N grows exponentially (e.g., for six-way associations, $N = 2^6 = 64$), it is possible that "add-one" smoothing may improve the MF estimator.

There seems to be no easy way to theoretically analyze why the "add-one" rule improve the MLE estimator.

5.6 How Many Samples Are Sufficient?

The answer depends on the trade-off between computation (and storage) costs and estimation errors. For very infrequent words, we might afford to sample 100%. In general, whenever possible we should try to sample as close to 2% as possible. The empirical "2% rule" is based on the observation that the conditional variance in (39) is proportional to $\frac{D}{D_s} - 1$. Figure 9(a) plots the relative standard error, i.e., $\sqrt{D/D_s - 1}$, as a function of the corpus sampling rate, D_s/D, indicating that the "elbow" point is around 2%.

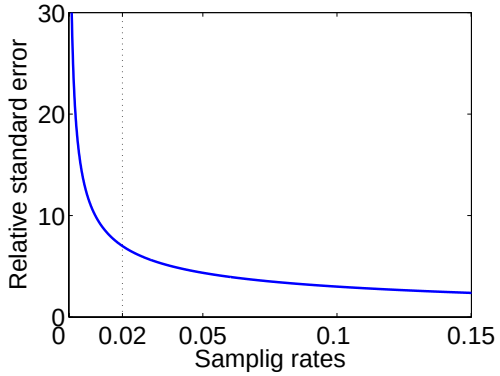


Figure 9
How large should the sampling rate be? We can sample up to the "elbow point" (2%), but after that there are diminishing returns (in terms of relative standard error reduction)

2% is certainly too large for high frequency words. At Web scale, 2% is also too large for "ordinary" words, whose document frequencies are in the order of 10 million. A more reasonable criterion is the coefficient of variation, cv = $\frac{\mathrm{SE}(\hat{a})}{a}$. We consider the estimate is accurate if the cv is below some threshold ρ_0 (e.g., 0.5). Note that the reciprocal of cv, $\frac{a}{\mathrm{SE}(\hat{a})}$ can be considered as the "Signal-to-Noise-Ratio."

The coefficient of variation can be expressed as

$$\mathrm{cv} = \frac{\mathrm{SE}(\hat{a})}{a} = \frac{1}{a}\sqrt{\frac{\frac{D}{D_s} - 1}{\frac{1}{a} + \frac{1}{f_1 - a} + \frac{1}{f_2 - a} + \frac{1}{D - f_1 - f_2 + a}}} \approx \frac{1}{a}\sqrt{\frac{\max\left(\frac{f_1}{k_1}, \frac{f_2}{k_2}\right) - 1}{\frac{1}{a} + \frac{1}{f_1 - a} + \frac{1}{f_2 - a} + \frac{1}{D - f_1 - f_2 + a}}}. \tag{59}$$

Figure 10(a) plots the required sampling rate $\min\left(\frac{k_1}{f_1}, \frac{k_2}{f_2}\right)$computed from (59). We assume $f_1 = \alpha_1 D$, $f_2 = \alpha_2 D$, $\alpha_2 \leq \alpha_1$, $a = \beta f_2$. In Figure 10(a), we have $\alpha_2 = 0.1\alpha_1$, $\beta = 0.05$, $\rho_0 = 0.5$, and for three different values of α_1 = 0.01, 0.001, 0.0001 $\times D$. The figure shows that at Web scales (i.e., $D \approx 10$ billion), a sampling rate as low as 10^{-4} may suffice for "ordinary" words (i.e., $f_1 \approx 10^7 = 0.001D$).

Figure 10(b) plots the required sample size k_1, for the same experiment in Figure 10(a). For simplicity, we assume $\frac{k_1}{f_1} = \frac{k_2}{f_2}$. The figure shows that, after D is large enough, the required sample size does not increase as much.

For better insights, we can simplify (59) by

$$\mathrm{cv} = \frac{\mathrm{SE}(\hat{a})}{a} \leq \frac{1}{a}\sqrt{\frac{\frac{D}{D_s}}{\frac{1}{a}}} = \sqrt{\frac{1}{a}\frac{D}{D_s}}. \tag{60}$$

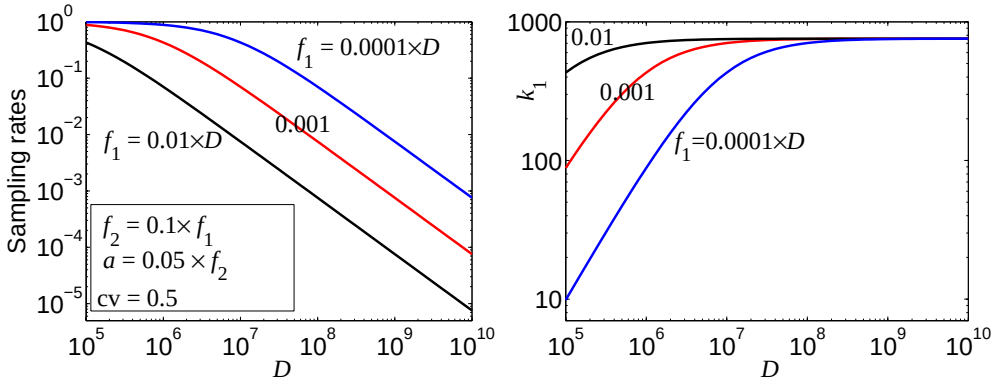


Figure 10
(a): An analysis based on cv = $\frac{\text{SE}}{a}$ = 0.5 suggests that we can get away with much lower sampling rates. The three curves plot the critical value for the sampling rate, $\frac{D_s}{D} \approx \min\left(\frac{k_1}{f_1}, \frac{k_2}{f_2}\right)$, as a function of corpus size, D. At Web scale, $D \approx 10^{10}$, sampling rates above 10^{-3} to 10^{-5} satisfy cv < 0.5, at least for these settings of f_x, f_y and a. The settings were chosen to simulate "ordinary" words. The three curves correspond to three choices of f_1: $D/100$, $D/1000$, and $D/10,000$. $f_2 = f_1/10$, $a = f_2/20$. (b) plots the critical sample size k_1 (assuming $\frac{k_1}{f_1} = \frac{k_2}{f_2}$), corresponding to the sampling rates in (a).

Suppose we would like cv $\leq \rho_0$, it suffices if $\sqrt{\frac{1}{a}\frac{D}{D_s}} \leq \rho_0$, i.e,

$$\frac{D_s}{D} \approx \min\left(\frac{k_1}{f_1}, \frac{k_2}{f_2}\right) \geq \frac{1}{\rho_0^2 a}, \tag{61}$$

which suggests that the sampling rate may decrease as the corpus scales up, because in general, we expect a increases with increasing D.

Assume $\frac{k_1}{f_1} = \frac{k_2}{f_2}$. In order for cv $\leq \rho_0$, it suffices if

$$\frac{k_1}{f_1} = \frac{1}{\rho_0^2 a} \Rightarrow k_1 = \frac{f_1}{\rho_0^2 a} = \frac{\alpha_1 D}{\rho_0^2 \beta \alpha_2 D} = \frac{\alpha_1}{\rho_0^2 \beta \alpha_2}, \tag{62}$$

which (approximately) explains why in Figure 10(b), the required sample size k_1 reaches a plateau after the corpus size D is larger than a certain level.

To apply (59) to the real data, Table 4 presents the critical sampling rates and sample sizes for all pair-wise combinations of the four-word query "Governor, Schwarzenegger, Terminator, Austria." Here we assume the estimates in Table 3 are exact. The table verifies that only a very small sample may suffice to achieve a reasonable cv.

The analysis above provides a nice solution for a single pair of words with particular values for f_1, f_2, and a. If these values are "representative" for "ordinary" words, then this analysis produces a rough estimate of typical sampling rates. If we would like to choose sample sizes more carefully, for all words in the corpus, we will have to seek other alternatives.

In many situations, we could compute the maximum allowed total sample size, for example, based on the available memory. That is, $\sum_{i=1}^{N} k_i = T$, where N is the total number of words and T is the maximum allowed total samples. We could allocate T according to document frequencies f_j, i.e.,

$$k_j = \frac{f_j}{\sum_{i=1}^{N} f_i} T. \tag{63}$$

Table 4
The critical sampling rates and sample sizes (for cv = 0.5) are computed for all two-way combinations among the four words "Governor, Schwarzenegger, Terminator, Austria," assuming the estimated document frequencies and two-way associations in Table 3 are exact. The required sampling rates are all very small, verifying our claim that for "ordinary" words, a sampling rate as low as 10^{-4} may suffice. In these computations, we used $D = 5 \times 10^9$ for the number of documents in the collection.

Query	Critical Sampling Rate	Sample Sizes
Governor & Schwarzenegger	2.2×10^{-6}	83 & 9
Governor & Terminator	2.9×10^{-5}	1084 & 101
Governor & Austria	5.5×10^{-6}	205 & 485
Schwarzenegger & Terminator	6.1×10^{-6}	25 & 21
Schwarzenegger & Austria	3.3×10^{-5}	131 & 2849
Terminator & Austria	2.2×10^{-5}	76 & 193

Usually, we will need to define a lowerbound $\underline{k}$ and an upperbound $\bar{k}$, which have to be selected from engineering experience, depending on the specific applications. We will truncate the computed k_j if it is outside $[\underline{k},\ \bar{k}]$. (63) implies a uniform sampling rate, which may not be always desirable, but the confinement by $[\underline{k},\ \bar{k}]$ can effectively vary the sampling rates.

Another reasonable criterion is to minimize the total number of "unused" samples. For a pair consists of word W_i and W_j, if $\frac{k_i}{f_i} \geq \frac{k_j}{f_j}$, then on average, there are $\left(\frac{k_i}{f_i} - \frac{k_j}{f_j}\right) f_i$ samples unused in K_i. This is the basic idea behind the following linear program for choosing the "optimal" sample sizes:

$$\begin{aligned} \text{Minimize} \quad & \sum_{i=1}^{N} \sum_{j=i+1}^{N} \left[f_i \left(\frac{k_i}{f_i} - \frac{k_j}{f_j} \right)_+ + f_j \left(\frac{k_j}{f_j} - \frac{k_i}{f_i} \right)_+ \right], \\ \text{subject to} \quad & \sum_{i=1}^{N} k_i = T, \qquad k_i \leq f_i, \qquad \underline{k} \leq k_i \leq \overline{k}, \end{aligned} \tag{64}$$

where $(z)_+ = \max(0, z)$, is the positive part of z.

This linear program can be modified easily to consider other factors in different applications. For example, some applications care more about the very rare words, so we would weight the rare words more.

So far, we have assumed that a is known, in computing the variance and cv. When a is unknown, we have to use the estimate a to estimate the variance. This is the typical situation in Sequential analysis (Siegmund, 1985). Using the results of (Chow and Robbins, 1965; Nadas, 1969), we can propose the following stopping rule:

$$D_s^* = \min \left\{ D_s \text{ such that } \frac{\text{SE}(\hat{a})}{\hat{a}} \leq \frac{\rho}{z_{1-\alpha/2}} \right\} \tag{65}$$

where $z_{1-\alpha/2}$ is the $\frac{1}{2}(1-\alpha)$ quantile of the standard Normal distribution. A common choice is $\alpha = 0.05$, $z_{1-\alpha/2} = 1.96 \approx 2$. ρ is the "proportional accuracy," which is defined as

$$|\hat{a} - a| \leq \rho a. \tag{66}$$

With this stopping rule, we can develop a sequential sampling scheme, in Algorithm 1.

Sequential sampling is particularly useful when the experiments are expensive or destructive (e.g., clinical trials). In our applications, if the postings are already in the memory, sequential

Algorithm 1 Sequential sampling algorithm

1: Choose α, ρ. For example, $\alpha = 0.05$, $\rho = 0.5$.
2: Choose a reasonable initial sample size, k_1, k_2.
3: Construct sketches K_1 and K_2. Estimate a. Use the estimated $\hat{a}$ to estimate the variance (the observed variance is recommended).
4: **if** $\frac{\mathrm{SE}(\hat{a})}{\hat{a}} \leq \frac{\rho}{z_{1-\alpha/2}}$ **then**
5: Exit
6: **else**
7: Increase k_1, k_2.
8: Goto 3
9: **end if**

sampling could be useful in saving the CPU time for estimating associations online. However, if the samples are in the memory but the original postings are stored on disk, sequential sampling could be very expensive. In this case, in the implementation of sequential sampling, we should try to read as many data as possible in one disk I/O.

5.7 When Will Sketch Not Work Well?

We consider three scenarios. (A): f_1 and f_2 are both large; (B): f_1 and f_2 are both small; (C): f_1 is very large but f_2 is very small. Conventional sampling over documents can handle situation (A), but will perform poorly on (B) because the less frequent words have less chance to be sampled (hence, large variance). The sketch algorithm can handle both (A) and (B) well. In fact, it will do very well when both words are small because the equivalent sampling rate $\frac{D_s}{D} \approx \min\left(\frac{k_1}{f_1}, \frac{k_2}{f_2}\right)$ can be very high, even 100%.

When $f_2 << f_1$, no sampling method can work well unless we are willing to sample P_1 with a sufficiently large sample. Otherwise even if we let $\frac{k_2}{f_2} = 100\%$, the corpus sampling rate, $\frac{D_s}{D} \approx \frac{k_1}{f_1}$, will be low.

For example, Google estimates 14,000,000 hits for "Holmes," 37,500 hits for "Diaconis," and 892 joint hits. [3] Assuming $D = 5 \times 10^9$ and cv = 0.5, the critical sample size for "Holmes" would have to be 6.1×10^4, probably way too large.

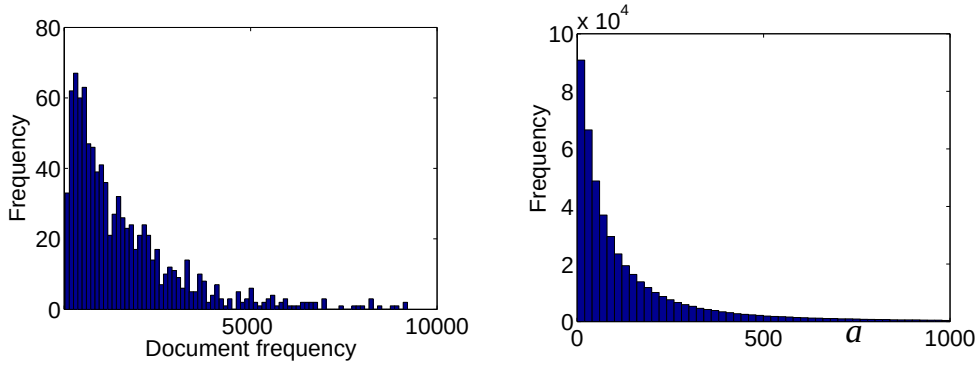
6 Evaluation of Two-way Associations

We evaluated our two-way association sampling/estimation algorithm with one chunk of web crawls provided by Microsoft MSN. The collection includes $D = 2^{16}$ web pages. We randomly sampled the English words that appeared in at least 20 documents and generated a dataset of 968 unique words (i.e., 468,028 pairs). Figure 11 displays the histograms and some statistics of the dataset.

The first (small dataset) experiment considered 4 English words (6 word pairs), shown in Table 5. The document IDs (from 1 to $D = 2^{16}$) were randomly permuted 10^5 times. On each permutation, we constructed sketches by sampling the postings at a range of sampling rates, from 0.002 to 0.95. With 10^5 Monte Carlo experiments, we are able to compute the mean square errors and other statistics to verify the correctness of our theoretical formulas (e.g., theoretical variance) and evaluate the performance of the various estimation methods we have studied.

After we have verified our theoretical formulas by the small dataset Monte Carlo experiment, we will run our algorithm on all 468,028 word pairs. In this larger dataset experiment, we will

3 As a married couple, Holmes and Diaconis are both professors in Statistics at Stanford University.



(a) (b)

Figure 11
(a): Histogram of the document frequencies (df) of the 968 words. Max df = 42564, median = 1135, mean = 2135, standard deviation = 3628. (b): Histogram of the co-occurrences (a) for the 468028 word pairs. Max a = 33045, mean = 188, median = 74, standard deviation = 459.

Table 5
Gold Standard associations, a. The document frequencies are shown in parentheses. These words are frequent, suitable for evaluating our algorithms at very low sampling rates. Since the associations are symmetric, they are only displayed in the lower triangle in the table. In the upper triangle, the six different combinations of word-pairs are numbered in square brackets.

	THIS	HAVE	HELP	PROGRAM
THIS (27633)	—	[2-1]	[2-2]	[2-3]
HAVE (17396)	13517	—	[2-4]	[2-5]
HELP (10791)	7221	5781	—	[2-6]
PROGRAM (5327)	3682	3029	1949	—

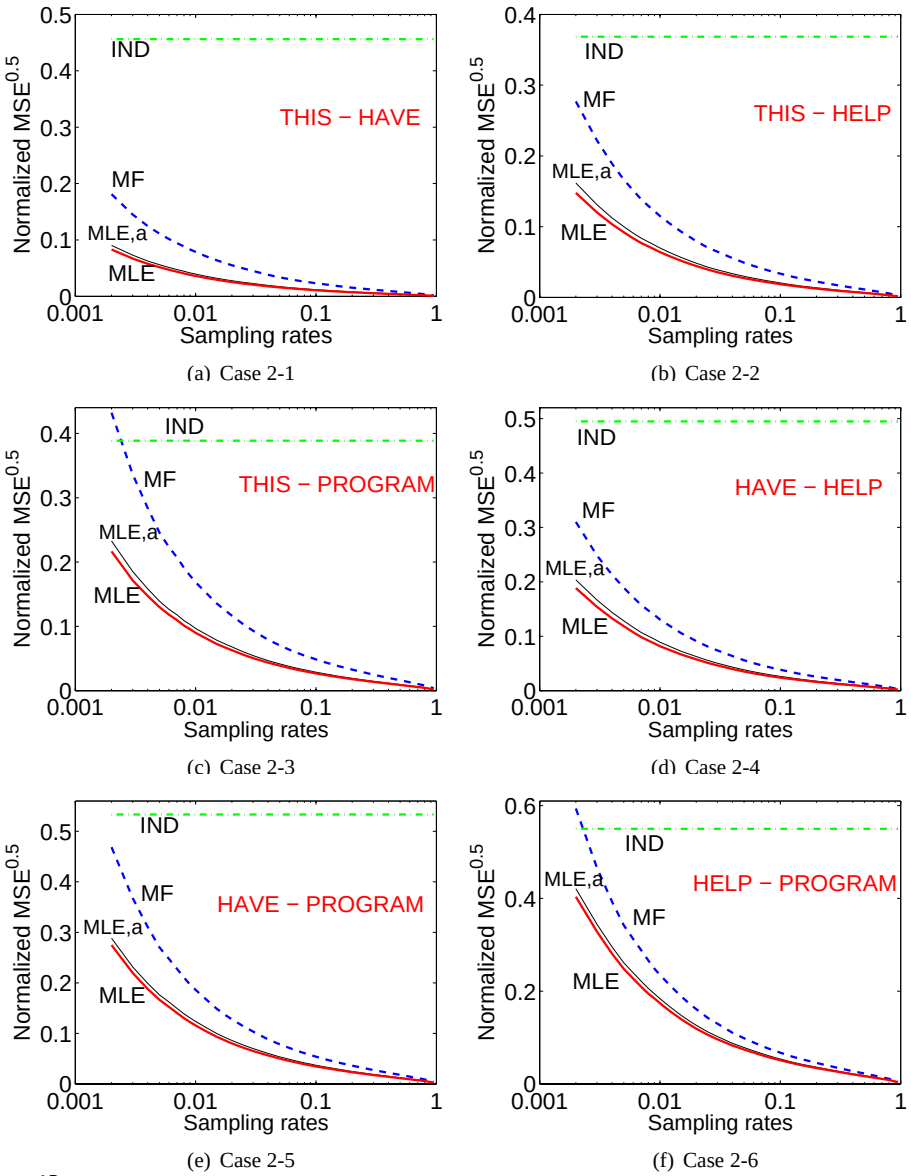
compare the performance of different estimators at a range of sampling rates. We repeat the experiment 6 times using different permutations.

6.1 Small Dataset Monte Carlo Experiment

Figure 12 evaluates the various estimate methods by MSE over a wide range of sampling rates. The figure shows that the proposed method, $\hat{a}_{MLE}$, is considerably better (by $20\% - 40\%$) than the margin-free baseline, $\hat{a}_{MF}$. The recommended approximation, $\hat{a}_{MLE,a}$, is remarkably close to the exact solution.

Figure 13 compares how the "add-one" smoothing affects the estimations ($\hat{a}_{MLE}$, $\hat{a}_{MLE,a}$, and $\hat{a}_{MF}$). The results are presented in terms of the percentages of improvements of MSE, with respect to the un-smoothed estimations. For all 6 cases, smoothing improves the proposed estimators $\hat{a}_{MLE}$ and $\hat{a}_{MLE,a}$, in some cases for up to 20% at low sampling rates. However, as expected, for all cases except Case 2-1, the "add-one" smoothing does not improve $\hat{a}_{MF}$. In fact, for Case 2-5 and Case 2-6, $\hat{a}_{MF+S}$ is $> 10\%$ worse than the $\hat{a}_{MF}$ at low sampling rates.

Figure 14 compares the theoretical unconditional variances with the empirical variances for two selected cases. We could use the approximate unconditional variance formula (54), which replaced E $\left(\frac{D}{D_s}\right)$ with its approximation $\max\left(\frac{f_1}{k_1}, \frac{f_2}{k_2}\right)$, but we decided to present the unconditional variances (42) using the measured (i.e., empirical) E $\left(\frac{D}{D_s}\right)$. Figure 15 plots the ratio of



(a) Case 2-1 (b) Case 2-2

(c) Case 2-3 (d) Case 2-4

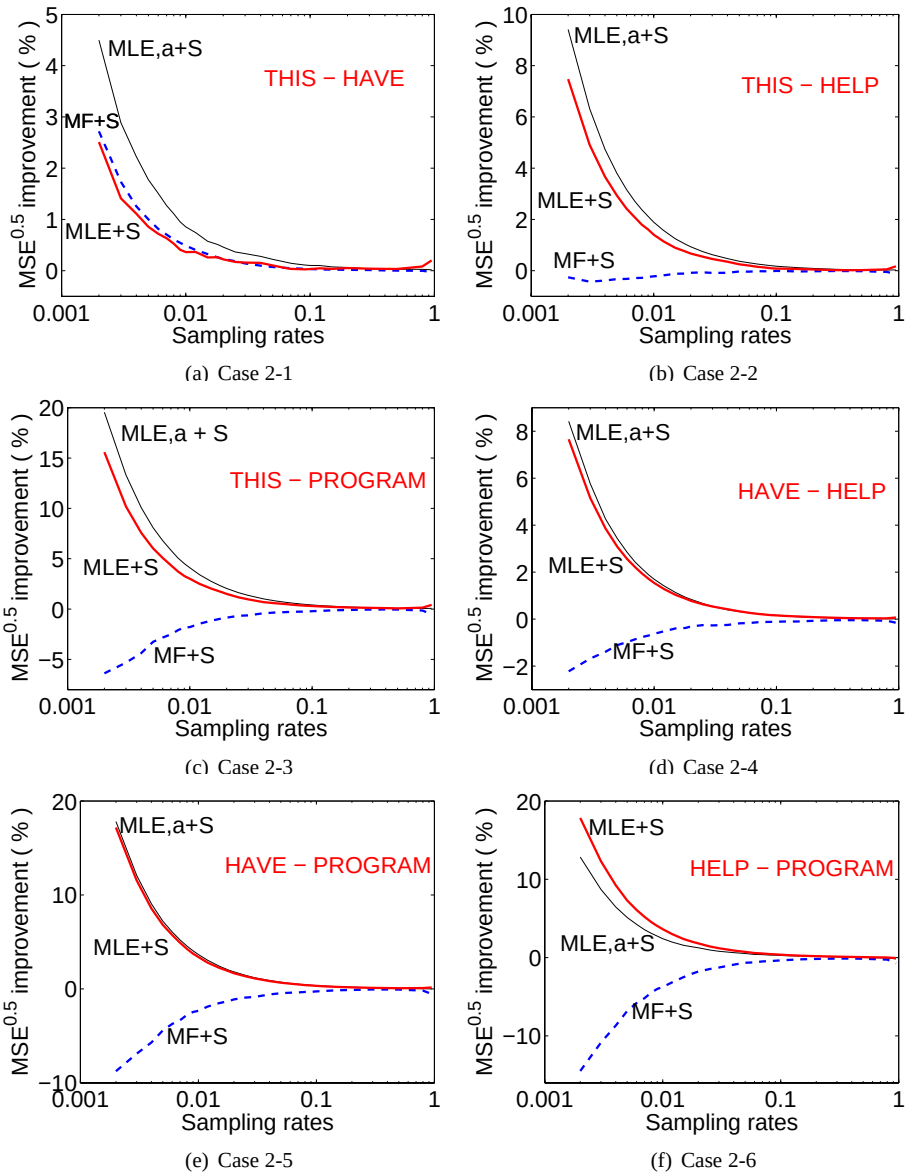
(e) Case 2-5 (f) Case 2-6

Figure 12

The proposed method, $\hat{a}_{MLE}$ outperforms the margin-free baseline, $\hat{a}_{MF}$, in terms of $\frac{\text{MSE}^{0.5}}{a}$. The recommended approximation, $\hat{a}_{MLE,a}$, is close to $\hat{a}_{MLE}$. All methods are better than assuming independence (IND).

$\max\left(\frac{f_1}{k_1}, \frac{f_2}{k_2}\right)$ over $\text{E}\left(\frac{D}{D_s}\right)$, to illustrate how much the approximate formula (54) underestimate the true unconditional variance.

Figure 14 indicates that, for the exact MLE, the theoretical variances match the empirical variances remarkable well. The figure also shows that the "add-one" smoothing is quite effective in reducing the variances for the exact MLE. For the margin-free (MF) estimator, the "add-one" smoothing does not reduce the variances (at least not noticeable). Also, as expected, the theoretical unconditional variance for the MF estimator slightly underestimates the true variances.



(a) Case 2-1 (b) Case 2-2

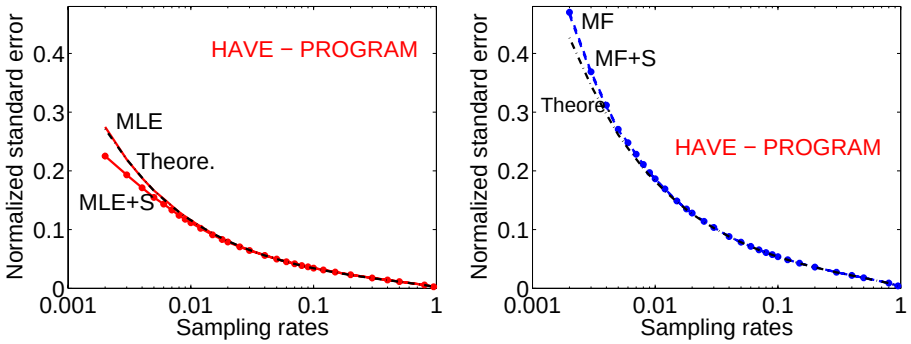
(c) Case 2-3 (d) Case 2-4

(e) Case 2-5 (f) Case 2-6

Figure 13
Smoothing improves the proposed MLE estimators but hurt the margin-free estimator in most cases. The vertical axis is the percentage of relative improvement in $\text{MSE}^{0.5}$ of each smoothed estimator with respect to its un-smoothed version.

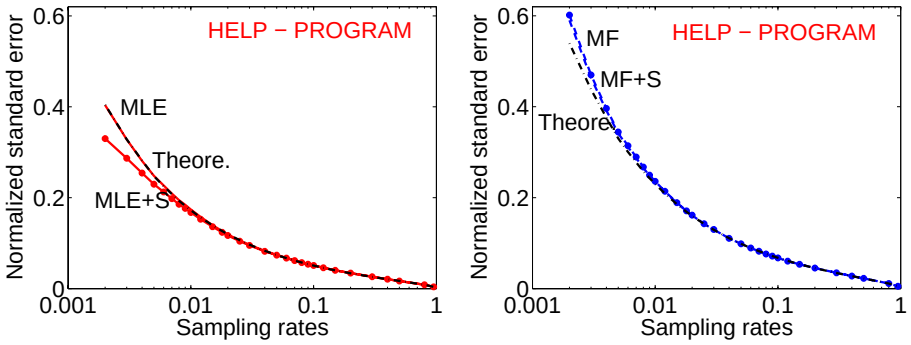
Figure 15 plots $\max\left(\frac{f_1}{k_1}, \frac{f_2}{k_2}\right) \Big/ \text{E}\left(\frac{D}{D_s}\right)$. This figure verifies that $\max\left(\frac{f_1}{k_1}, \frac{f_2}{k_2}\right) \leq \text{E}\left(\frac{D}{D_s}\right)$ but the differences are not very large. For example, at a sampling rate of 0.01, for all 6 cases, $\max\left(\frac{f_1}{k_1}, \frac{f_2}{k_2}\right) \geq 0.95\text{E}\left(\frac{D}{D_s}\right)$. Therefore, while it appears difficult to compute $\text{E}\left(\frac{D}{D_s}\right)$, it is fairly accurate to use the simple approximation: $\max\left(\frac{f_1}{k_1}, \frac{f_2}{k_2}\right)$.

Finally, we also compare the biases in Figure 16 for Case 2-5 and Case 2-6. The figure shows that the MLE estimator is essentially unbiased, unlike the margin-free baseline.



(a) Case 2-5, MLE

(b) Case 2-5, MF



(c) Case 2-6, MLE

(d) Case 2-6, MF

Figure 14

Normalized standard error, $\frac{\mathrm{SE}(\hat{a})}{a}$. For the MLE, the theoretical unconditional variance formula (42) fits the simulation results so well that the curves are indistinguishable. Also, smoothing effectively reduces the variances at low sampling rates. In contrast, the margin-free estimator exhibits higher variances than the MLE and smoothing does not reduce variances. In addition, the theoretical variance for the MF estimator under-estimates the true variance.

6.2 Large Dataset Experiment

The large experiment considers 968 English words (468,028 pairs) over a range of sampling rates, from 0.003 to 0.5. A floor of sampling rates is imposed so no sample contains fewer than 20 documents.

As reported in Figure 17, the large experiment confirms again that the proposed method, $\hat{a}_{MLE}$, is considerably better than the margin-free baseline, which is also better than the independence baseline. The recommended approximation, $\hat{a}_{MLE,a}$, is very close to $\hat{a}_{MLE}$.

Figure 18(a) plots the percentage of improvements after applying the "add-one" smoothing. For the MLE method, smoothing can improve as much as 20%. For the margin-free method, smoothing worsens the performance. Figure 18(b) more clearly displays how much $\hat{a}_{MLE+S}$ does better than $\hat{a}_{MF}$. The differences are, in general, about $20\% - 30\%$.

Therefore, both small dataset and large dataset experiments verify that the proposed MLE method considerably improves the margin-free method. The approximate MLE method, which, like the margin-free method, has a closed-form solution, produces remarkably close results to the exact MLE. The simple "add-one" smoothing helps the both proposed MLE estimators but hurts the margin-free method.

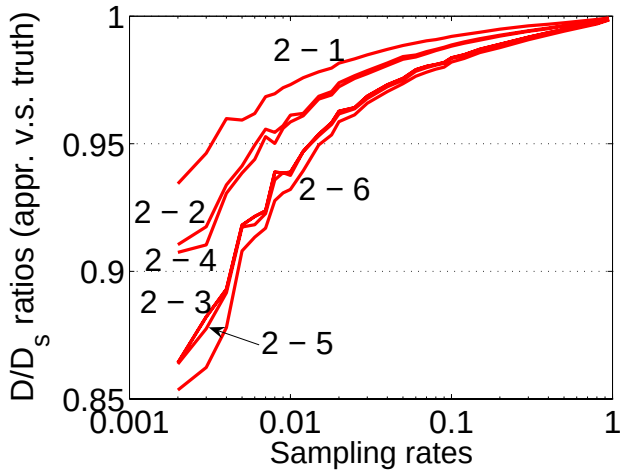
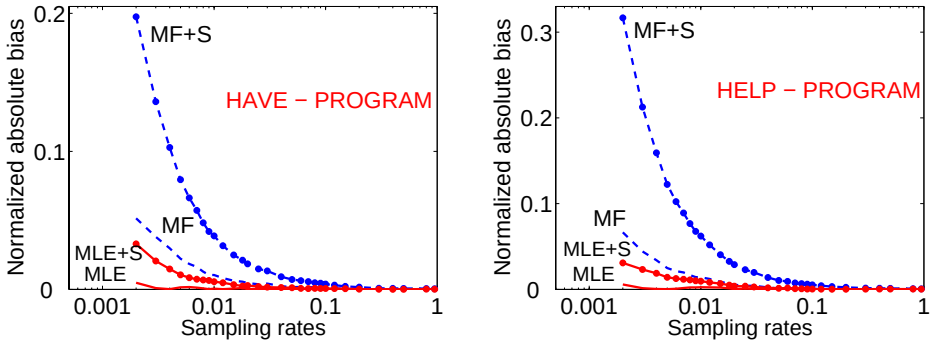


Figure 15

For all 6 cases, the ratios $\max\left(\frac{f_1}{k_1}, \frac{f_2}{k_2}\right) \Big/ \mathrm{E}\left(\frac{D}{D_s}\right)$ are close to 1, and the differences roughly monotonically decrease. When the sampling rates $>= 0.01$, $\max\left(\frac{f_1}{k_1}, \frac{f_2}{k_2}\right)$ is a very accurate approximation of $\mathrm{E}\left(\frac{D}{D_s}\right)$.



(a) Case 2-5 (b) Case 2-6

Figure 16

Biases in terms of $\frac{|\mathrm{E}(\hat{a})-a|}{a}$. $\hat{a}_{MLE}$ is practically unbiased, unlike $\hat{a}_{MF}$. Smoothing increases bias slightly.

6.3 Rank Retrieval by Cosine Similarity

These estimates of a can be used to find highly associated pairs of words using standard similarity measures such as cosine: $\frac{a}{\sqrt{f_1 f_2}}$. Because the margins f_1 and f_2 are known, estimating the cosine coefficient is equivalent to estimating the joint frequency, a. If we sort word pairs by their cosines, using estimates of a based on a small sample, the rankings will hopefully be close to what we would obtain if we used all the data. This section will compare the rankings based on a small sample to a gold standard, the rankings based on all of the data.

How should we evaluate the two rankings? One simple measure is "top-k" percentage of agreements. That is, we compare the top-k pairs from the reconstructed list with the top-k list from the gold standard list and compute the percentage of how many pairs are common in both "top-k" lists, as in (Ravichandran et al., 2005).

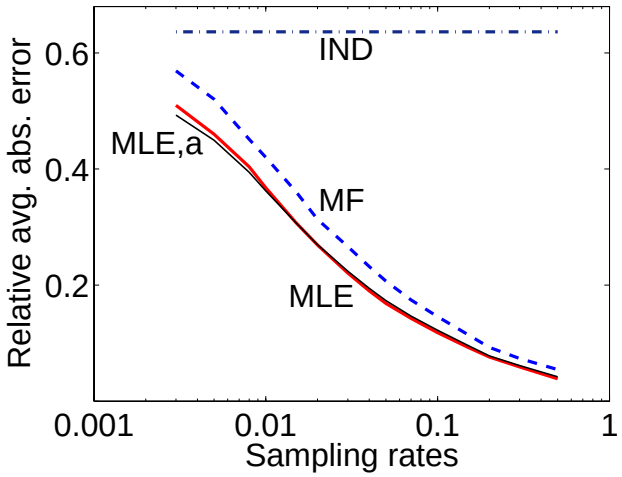
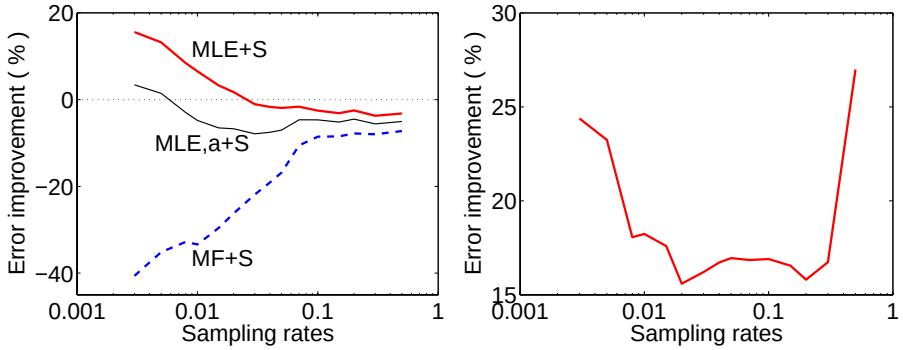


Figure 17
We report the (normalized) mean absolute errors (divided by the mean co-occurrences, 188). All curves are averaged over three permutations. The proposed MLE and the recommended approximation are very close and both are significantly better than the margin-free (MF) baseline. All estimators do better than assuming independence.



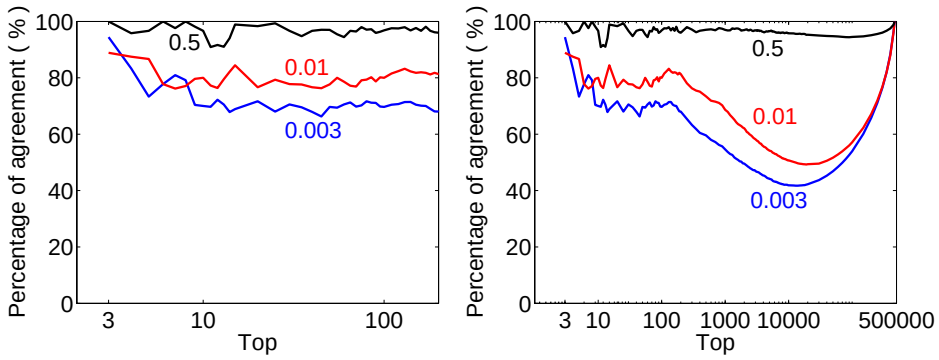
(a) Smoothing (b) MF v.s. MLE+S

Figure 18
(a): The "add-one" smoothing improves the proposed (exact and approximate) MLE methods, but hurts the margin-free method. (b): $\hat{a}_{MLE+S}$ improves $\hat{a}_{MF}$ for $15\% - 30\%$ at most sampling rates.

Figure 19(a) plots the top-3 to top-200 percentage of agreements for the 468,028 pairs in the large dataset experiment. With a sampling rate of 0.003, the agreements are consistently around 70%. With a sampling rate of 0.5, the agreements are close to 100%. Increasing sampling rates, increases agreements.

Figure 19(b) presents the same type of results as in Figure 19(a), but not limited to top-200. The figure suggests that the most strongly-associated (e.g., top-200 in this case) pairs are likely to remain in the top of the similarity list even with very low sampling rates.

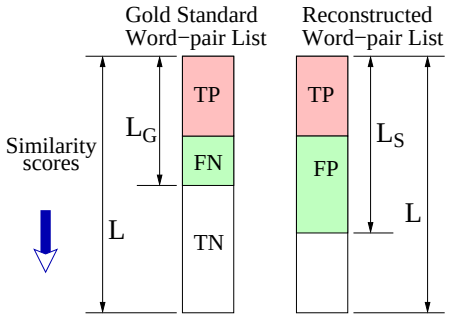
The same comparisons can be evaluated in terms of precision and recall, by fixing the top-L_G gold standard list but varying the length of the sample list L_S. More precisely, recall $=$ relavant$/L_G$, and precision $=$ relevant$/L_S$, where "relevant" means the retrieved pairs in the gold standard list. Figure 20 gives a graphical representation of this evaluation scheme, using



(a) Top 3 - 200 (b) Top 3 - All

Figure 19
The 468,028 pairs were sorted descending by the cosine similarity scores to form a similarity list. The gold standard similarity list was constructed using the true associations, while the reconstructed similarity list used the estimated a. The vertical axis is the percentage of agreements among the top-k word pairs between the two similarity lists. (a): k ranges from 3 to 200. (b): k ranges from 3 to 468,028. In both sub-figures, there are three curves, corresponding to three different sampling rates: 0.003, 0.01, and 0.5.

notation in (Manning and Schutze, 1999, Chapter 8.1): true positive (TP) = retrieved and relevant, false negative (FN) = relevant but not retrieved, false positive (FP) = retrieved and irrelevant, true negative (TN) = irrelevant and not retrieved.



	Relevant	Irrelevant
Retrieved	TP (a)	FP (b)
Not Retrieved	FN (c)	TN (d)

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{a}{a + b}$$

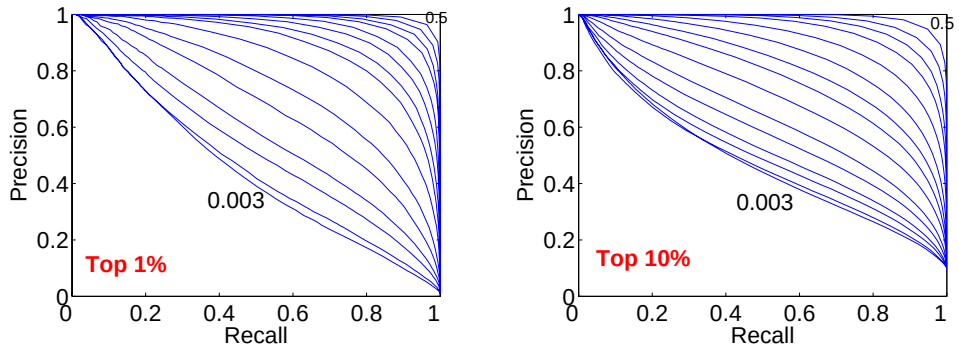
$$\text{Recall} = \frac{TP}{TP + FN} = \frac{a}{a + c}$$

Figure 20
Definitions of recall and precision. L = total number of pairs. L_G = number of pairs from the top of the gold standard similarity list. L_S = number of pairs from the top of the reconstructed similarity list.

Figure 21 presents the precision-recall curves for $L_G = 1\%L$ and $10\%L$, where $L = 468,028$. For each L_G, there is one precision-recall curve corresponding to each sampling rate. All curves indicate the precision-recall trade-off and that the only way to improve both precision and recall simultaneously is to increase the sampling rate.

7 Estimate Two-way Contingency Table Summary Statistics

After we have estimated a, we can then compute all summary statistics, such as the cosine coefficient, dice coefficient, Jaccard coefficient, $\log$ likelihood ratio (LLR), generalized inverse document frequency (IDF), and more. We denote these summary statistics by $h(a)$ generically. The simplest estimator of $h(a)$ would be $h(\hat{a})$, i.e, substituting the estimated a into $h(a)$. However, this method may raise some concerns, especially when $h(a)$ is a strong non-linear function



(a) Top 1% (b) Top 10%

Figure 21
Precision-recall curves in retrieving the top-1% and top-10% gold standard pairs, at different sampling rates from 0.003 to 0.5. Note that the precision is always larger than $\frac{L_G}{L}$.

of a and we have small samples.

We have shown that our proposed MLE method is practically unbiased. However, $h(\hat{a})$ in general is not unbiased unless $h(a)$ is a linear function of a. If $h(a)$ is a convex function of a, then by Jensen's inequality, $\mathrm{E}(h(a)) \geq h(\mathrm{a})$. Similarly, if $h(a)$ is concave, then $\mathrm{E}(h(a)) \leq h(\mathrm{a})$. See Figure 22 for the example of LLR. We should mention that, when the first derivative $h'(a)$ exists and non-zero, $h(a)$ is in fact asymptotically unbiased, but the convergence rate could be very slow (depending on h).

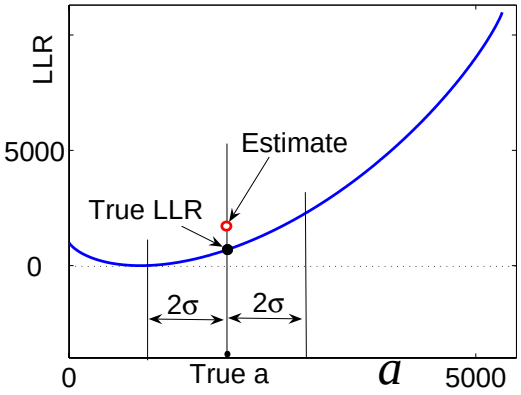


Figure 22
The log likelihood ratio (LLR) curve is plotted for Case 2-6 in Table 5, i.e., W_1 = "HELP" with $f_1 = 10791$, W_2 = "PROGRAM" with $f_2 = 5327$, $a = 1949$, $D = 2^{16}$. LLR is a convex function of a. If we use the estimated a to compute LLR, i.e., LLR$(\hat{a})$, due to the errors (variance) of $\hat{a}$ and the convexity of LLR, LLR$(\hat{a})$ will be above the LLR curve.

If it is desirable to have an unbiased estimator of $h(a)$, we can correct the biased estimator $h(\hat{a})$ by a Taylor expansion:

$$h(\hat{a}) = h(a) + (\hat{a} - a)h'(a) + \frac{(\hat{a} - a)^2}{2}h''(a) + \frac{(\hat{a} - a)^3}{6}h'''(a) + \text{negligible terms}, \quad (67)$$

where $h'(a)$, $h''(a)$, and $h'''(a)$, are the first, second, and third derivatives, respectively. Taking

expectations in both sides, we obtain

$$\mathrm{E}(h(\hat{a})) \approx h(a) + \frac{\sigma^2}{2}h''(a) + \frac{\eta^3}{6}h'''(a), \tag{68}$$

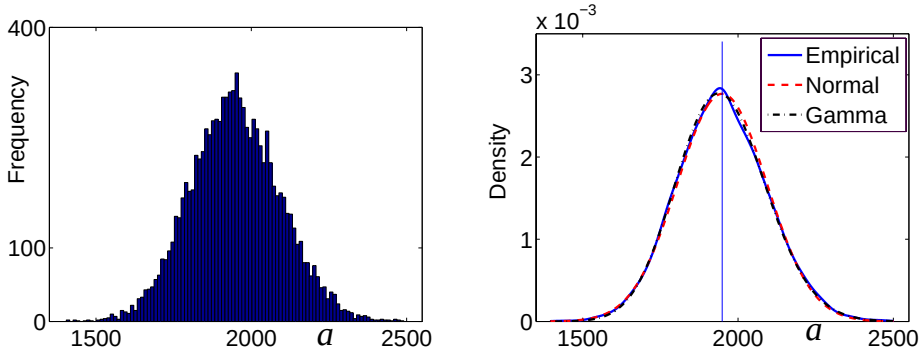
where $\sigma^2 = \mathrm{Var}(\hat{a}_{MLE})$, and $\eta^3 = \mathrm{E}\left(\hat{a}_{MLE} - a\right)^3$. Here we treat $\hat{a}_{MLE}$ as an unbiased estimator, i.e., $\mathrm{E}(\hat{a}_{MLE} - a) \approx 0$.

Equation (68) suggests an approximately unbiased estimator, denoted by $\hat{h}(a)$,

$$\hat{h}(a) = h(\hat{a}) - \frac{\sigma^2}{2}h''(a) - \frac{\eta^3}{6}h'''(a). \tag{69}$$

We already know how to compute σ^2, at least approximately. The large sample theory says that $\hat{a}_{MLE}$ can be approximated as a Normal random variable, i.e., $\hat{a} \sim \mathrm{N}\left(a, \sigma^2\right)$. Under the Normal assumption, all odd central moments of $\hat{a}$ vanish (e.g., $\eta^3 = 0$). We refer to this adjustment as the "Normal correction."

As mentioned in Section 5.3, we could use another (asymptotically) equivalent distribution to approximate the distribution of $\hat{a}_{MLE}$, for example, a Gamma random variable, $\mathrm{G}(\alpha, \beta)$, as long as the first two moments are the same. If we replace Normal with Gamma, we refer to the adjustment as "Gamma Correction." Figure 23 gives an example of the histogram and empirical density of $\hat{a}_{MLE}$, which shows that both Normal and Gamma fit the empirical data quite well.



(a) Histogram (b) Histogram

Figure 23
(a) The histograms of $\hat{a}_{MLE}$ for Case 2-6 in Section 6 at a sampling rate 0.05. Recall that Case 2-6 involved word "HELP" and word "PROGRAM" with true association $a = 1949$. (b): The empirical density of $\hat{a}_{MLE}$ is very close to Normal, which verifies the large sample theory. The asymptotically equivalent Gamma approximation also fits the empirical density well.

The first two moments of $\mathrm{N}(a, \sigma^2)$ are a and σ^2, respectively. The first two moments of $\mathrm{G}(\alpha, \beta)$ are $\alpha\beta$ and $\alpha\beta^2$, respectively (Shao, 1999, Table 1.2). Equating the first two moments of Normal and Gamma, we obtain,

$$\alpha = \frac{a^2}{\sigma^2}, \qquad \beta = \frac{\sigma^2}{a}. \tag{70}$$

Assuming Gamma $\mathrm{G}(\alpha, \beta)$, the third central moment η^3 would be

$$\eta^3 = 2\alpha\beta^3 = 2\frac{\sigma^4}{a}. \tag{71}$$

This technique for removing biases is commonly used in statistics, see (Lehmann and Casella, 1998, Theorem 6.1.1) for an example. Often in practice, the Taylor expansion of $h(\hat{a})$ is truncated

after the second-derivative term. The reason that we recommend a third-order Taylor expansion and using a Gamma distribution to replace the Normal is because most of the summary statistics we are interested in are positive, while using only a second-order Taylor expansion may lead to negative values. We use the $\log$ likelihood ratio (LLR) to illustrate this effect. This "Gamma trick," i.e., replacing the Normal with a Gamma as the asymptotic distribution, was also used by Li et al. (2005).

In terms of D, f_1, f_2, and a, LLR, can be expressed as (Agresti, 2002, Section 3.2.1),

$$\begin{aligned}\text{LLR}(a) = a\log\frac{Da}{f_1 f_2} + (f_1-a)\log\frac{(f_1-a)D}{(D-f_2)f_1} + (f_2-a)\log\frac{(f_2-a)D}{(D-f_1)f_2} \\ + (D-f_1-f_2+a)\log\frac{(D-f_1-f_2+a)D}{(D-f_1)(D-f_2)}.\end{aligned} \tag{72}$$

The first three derivatives of LLR [4] are

$$\text{LLR}'(a) = \log\frac{a(D-f_1-f_2+a)}{(f_1-a)(f_2-a)}, \tag{73}$$

$$\text{LLR}''(a) = \frac{1}{a} + \frac{1}{f_1-a} + \frac{1}{f_2-a} + \frac{1}{D-f_1-f_2+a}, \tag{74}$$

$$\text{LLR}'''(a) = -\frac{1}{a^2} + \frac{1}{(f_1-a)^2} + \frac{1}{(f_2-a)^2} - \frac{1}{(D-f_1-f_2+a)^2}. \tag{75}$$

The second derivative, $\text{LLR}''(a)$, is positive, i.e., it is possible that the Normal correction, i.e., $\hat{\text{LLR}}(a) = \text{LLR}(\hat{a}) - \frac{\sigma^2}{2}\text{LLR}''(a)$, may be negative in some situations. Note that using a higher-order Taylor expansion would not avoid this problem because all even derivatives of LLR are positive.

Figure 24 compares the two bias correction methods. Although both methods do well in removing the bias in estimating LLR, the Normal correction generates so many negative values that we do not recommend it.

After we have adjusted the estimator for $h(a)$, then how about the variance?

A first-order Taylor expansion of $h(\hat{a})$ leads to:

$$\text{Var}(h(\hat{a})) \approx \sigma^2(h'(a))^2, \tag{76}$$

which is also well-known as the popular "Delta Method" (Agresti, 2002, Chapter 3.1.5). Note that (76) is asymptotically exact as long as $h(a)$ exists and non-zero, although the rate of convergence may be very slow.

If $h(a)$ is a convex function of a, then $h(\hat{a}) - h(a) \geq (\hat{a}-a)h'(a)$ and $\text{E}\left(h(\hat{a}) - h(a)\right) \geq \text{E}\left((\hat{a}-a)h'(a)\right)$, i.e., $\text{Var}(h(\hat{a})) \geq \sigma^2(h'(a))^2$. Similarly if $h(a)$ is a concave function of a, we know $\text{Var}(h(\hat{a})) \leq \sigma^2(h'(a))^2$.

Theoretically, the adjusted $\hat{h}(a)$ in (69) will have the same variance as $h(\hat{a})$ because σ^2 and η^3 are constants, i.e., $\text{Var}\left(\hat{h}(a)\right) = \text{Var}(h(\hat{a}))$. However, if we have to use the estimated a to compute σ^2 and η^3, $\hat{h}(a)$ should have larger variance than $h(\hat{a})$ because of additional variations.

To this end, we have discussed LLR in details. Table 6 collects some other common summary statistics including the angle (inverse cosine), the generalized IDF, and three vector similarity

4 The derivatives of LLR contain some other useful information. The first derivative, $\text{LLR}'(a)$ is the same as the "log odds ratio," an important summary statistics of contingency tables (Agresti, 2002, Chapter 2.2.3, Chapter 3.1.1). The zero first-derivative point of $\text{LLR}(a)$, i.e., $a = \frac{f_1 f_2}{D}$, corresponds to the point where the odds ratio = 1 ($\log$ odds ratio = 0). In other words, when two words are independent, LLR has zero first derivative. The second derivative of LLR is always positive, i.e., LLR is convex in a and reaches its minimum (= 0) at $a = \frac{f_1 f_2}{D}$, which is consistent with the well-known facts about mutual information, as LLR is basically the mutual information.

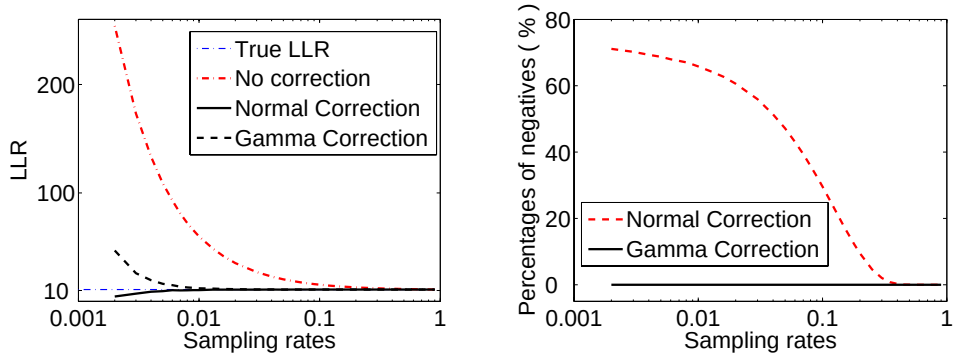


Figure 24
Both Normal and Gamma corrections were applied to the case: $f_1 = 10791$, $f_2 = 5327$, $a = 1000$, $D = 2^{16}$, which is basically Case 2-6 in Section 6 except that a was reduced from 1949 to 1000 to make the case more difficult. The true LLR = 10.86. 10^6 random samples were generated from N(a, σ^2), where σ^2 was the approximate unconditional variance (54). For each sample, an LLR value was computed. The average over all 10^6 LLR values is the uncorrected estimator of LLR(a) in (a). Both Normal correction and Gamma correction, on average, fit the exact LLR very well except at very small sampling rates. The Normal correction appears even better than the Gamma correction. The sub-figure (b) plots the percentage of negative values after corrections, among 10^6 samples. With the Normal correction, a large portion (can be 70% at low sampling rates) are negative, while with the Gamma correction, no negative values are observed in this example.

measures: dice, cosine, and Jaccard (resemblance), along with their first three moments. The dice and cosine coefficients are linear functions of a hence no need for them to be adjusted for bias. Resemblance is a weak non-linear function of a and therefore we expect its bias to be negligible. The second derivative of the generalized IDF is $\frac{1}{a^2}$, implying its non-linearity is not as strong as LLR, whose second derivative is $O\left(\frac{1}{a}\right)$. The non-linearity of the angle (inverse cosine) depends on f_1, f_2, and a. In most cases, we expect its second derivative to be quite small (i.e., small bias), but when $a \approx f_1 \approx f_2$ (i.e., angle ≈ 0), its non-linearity is severe. When $a = f_1 = f_2$, the first derivative becomes ∞ and we can no longer use the Taylor expansion.

The next two sections discuss the estimation of resemblance and cosine angle in more detail.

8 Estimate Resemblance from Contingency Tables

In this section, we will show that our proposed MLE method always outperforms Broder's sketch. The improvement is roughly a factor of 2 in normal settings.

Broder's sketch algorithm was originally designed to estimate the resemblance (Jaccard coefficient): $R(a) = \frac{a}{a+b+c} = \frac{a}{f_1+f_2-a}$. In this section, we will compute R from the estimated contingency table with:

$$\hat{R}_{MLE} = \frac{\hat{a}_{MLE}}{f_1 + f_2 - \hat{a}_{MLE}}. \tag{77}$$

$\hat{R}_{MLE}$ is slightly biased (see Table 6). We could correct the bias using a second-order or third-order Taylor expansion as in described in the previous section. However, since the second derivative of $R(a)$

$$R''(a) = \frac{2(f_1+f_2)}{(f_1+f_2-a)^3} \leq \frac{2(f_1+f_2)}{\max(f_1,f_2)^3} \leq \frac{4}{\max(f_1,f_2)^2}, \tag{78}$$

Table 6
Definitions and first three derivatives of some common summary statistics. The dice and cosine measures are linear functions of a. The second derivatives of resemblance is very small and the second derivative of the generalized IDF is on the order of $O\left(\frac{1}{a^2}\right)$, which is also small, hinting that bias corrections for resemblance or generalized IDF may not be as important as for LLR. The inverse cosine function has different degrees of non-linearity.

	$h(a)$	$h'(a)$	$h''(a)$	$h'''(a)$
Dice	$\frac{2a}{f_1+f_2}$	$\frac{2}{f_1+f_2}$	0	0
Cosine	$\frac{a}{\sqrt{f_1 f_2}}$	$\frac{1}{\sqrt{f_1 f_2}}$	0	0
Jaccard (Resemblance)	$\frac{a}{f_1+f_2-a}$	$\frac{f_1+f_2}{(f_1+f_2-a)^2}$	$\frac{2(f_1+f_2)}{(f_1+f_2-a)^3}$	$\frac{6(f_1+f_2)}{(f_1+f_2-a)^4}$
Generalized IDF	$\log(\frac{D}{a})$	$-\frac{1}{a}$	$\frac{1}{a^2}$	$-\frac{2}{a^3}$
Angle	$\cos^{-1}\left(\frac{a}{\sqrt{f_1 f_2}}\right)$	$\frac{-1}{\sqrt{f_1 f_2-a^2}}$	$\frac{-a}{(f_1 f_2-a^2)^{\frac{3}{2}}}$	$\frac{-1}{(f_1 f_2-a^2)^{\frac{3}{2}}}+\frac{-3a^2}{(f_1 f_2-a^2)^{\frac{5}{2}}}$

is very small, it is unlikely that the bias will have a noticeable effect.

The variance of $\hat{R}_{MLE}$ is approximately:

$$\begin{aligned}\text{Var}\left(\hat{R}_{MLE}\right) &\approx \text{Var}(\hat{a}_{MLE})(R'(a))^2 \\ &= \frac{\max\left(\frac{f_1}{k_1}, \frac{f_2}{k_2}\right)}{\frac{1}{a}+\frac{1}{f_1-a}+\frac{1}{f_2-a}+\frac{1}{D-f_1-f_2+a}} \frac{(f_1+f_2)^2}{(f_1+f_2-a)^4}, \end{aligned} \tag{79}$$

ignoring the "finite-population correction factor," for convenience.

Broder considered the special case where the two sample sizes were the same: $k_1 = k_2 = k$. In the original sketch construction (Broder, 1997), the estimator of resemblance, denoted as R_B, has a hypergeometric distribution. In the "minwise" sketch construction (Broder et al., 1998), the estimator, denoted as $R_{B,r}$, is a binomial. We have reviewed the Broder's original sketch in Section 2. For completeness, we shall also review the "minwise" sketch construction as follows.

After a random permutation on the document ID's, we record the smallest IDs in the postings P_1 and P_2, denoted as MIN(P_1) and MIN(P_2), respectively. The possibility of MIN(P_1) = MIN(P_2) would be $|\mathrm{P}_1 \cap \mathrm{P}_2|$ out of $|\mathrm{P}_1 \cup \mathrm{P}_2|$, i.e.,

$$P\left(\text{MIN}(\mathrm{P}_1) = \text{MIN}(\mathrm{P}_2)\right) = \frac{|\mathrm{P}_1 \cap \mathrm{P}_2|}{|\mathrm{P}_1 \cup \mathrm{P}_2|}, \tag{80}$$

$P\left(\text{MIN}(\mathrm{P}_1) = \text{MIN}(\mathrm{P}_2)\right)$ can be estimated by repeating the permutation k times independently.

The "minwise" construction appears to be more straightforward than Broder's original sketch. However, the original sketch used only one permutation while the "minwise" construction used k permutations. One of the reasons that Broder moved to the "minwise" construction is to overcome the difficulty in dealing with very short postings. Recall Broder assumed equal sample sizes, which can be problematic for very short postings because the pre-specified sample size has

to be large enough to ensure accuracy. This, however, is not a problem for our generalization of the sketch algorithm since we do not assume equal samples.

For simplicity, we ignore the difference between R_B and $R_{B,r}$. As a binomial, the variance would be

$$\text{Var}\left(\hat{R}_B\right) \approx \text{Var}\left(\hat{R}_{B,r}\right) = \frac{1}{k}R(1-R) = \frac{1}{k}\frac{a(f_1+f_2-2a)}{(f_1+f_2-a)^2}.$$

We can use the ratio $V_B = \frac{\text{Var}\left(\hat{R}_{MLE}\right)}{\text{Var}\left(\hat{R}_B\right)}$ to compare the performance of our proposed MLE with Broder's sketch:

$$V_B = \frac{\text{Var}\left(\hat{R}_{MLE}\right)}{\text{Var}\left(\hat{R}_B\right)} = \frac{\max\left(\frac{f_1}{k_1}, \frac{f_2}{k_2}\right)}{\frac{1}{a}+\frac{1}{f_1-a}+\frac{1}{f_2-a}+\frac{1}{D-f_1-f_2+a}} \frac{(f_1+f_2)^2}{(f_1+f_2-a)^2} \frac{k}{a(f_1+f_2-2a)}. \tag{81}$$

In most cases, it is reasonable to assume that $a << \min(f_1, f_2) < \max(f_1, f_2) << D$, i.e., $\text{Var}\left(\hat{R}_{MLE}\right) \approx \max(\frac{f_1}{k_1}, \frac{f_2}{k_2})\frac{a}{(f_1+f_2)^2}$, $\text{Var}\left(\hat{R}_B\right) \approx \frac{1}{k}\frac{a}{f_1+f_2}$. Therefore, approximately

$$V_B \approx \frac{k\max(\frac{f_1}{k_1}, \frac{f_2}{k_2})}{f_1+f_2}. \tag{82}$$

With equal samples, i.e., $k_1 = k_2 = k$, we have

$$V_B \approx \frac{\max(f_1, f_2)}{f_1+f_2}, \tag{83}$$

which is about $\frac{1}{2}$ when $f_1 = f_2$.

With proportional samples, i.e., $k_1 = 2k\frac{f_1}{f_1+f_2}$, $k_2 = 2k\frac{f_2}{f_1+f_2}$, we have

$$V_B \approx \frac{1}{2}. \tag{84}$$

As previously mentioned, in Broder's sketch construction, only half of the samples are used in the estimation. Our construction uses more samples. In fact, with proportional sampling, almost all samples will be used. These observations are consistent with V_B.

$V_B \approx \frac{1}{2}$ suggests that our algorithm is a significant improvement over Broder's original sketch. It implies that in order to achieve the same accuracy, our method requires only half as many samples as in Broder's construction.

Figure 25 plots V_B in (81) for the whole range of f_1, f_2, and a, assuming equal samples: $k_1 = k_2 = k$. We can see that $V_B \leq 1$ always holds and $V_B = 1$ only when $f_1 = f_2 = a$, which is a trivial case. When a/f_2 is small, $V_B \approx \frac{\max(f_1,f_2)}{f_1+f_2}$ holds well.

Compared with equal samples in Figure 25, proportional samples further reduce V_B, as shown in Figure 26.

It is not hard to show algebraically that V_B in (81) is always less than unity unless $f_1 = f_2 = a$. For convenience, we use the notion a, b, c, d in (81). Assume $k_1 = k_2 = k$ and $f_1 > f_2$, we obtain

$$V_B = \frac{a+b}{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}} \frac{(2a+b+c)^2}{(a+b+c)^2} \frac{1}{a(b+c)}. \tag{85}$$

To show $V_B \leq 1$, suffices to show

$$(a+b)(2a+b+c)^2bcd \leq (bcd+acd+abd+abc)(a+b+c)^2(b+c), \tag{86}$$

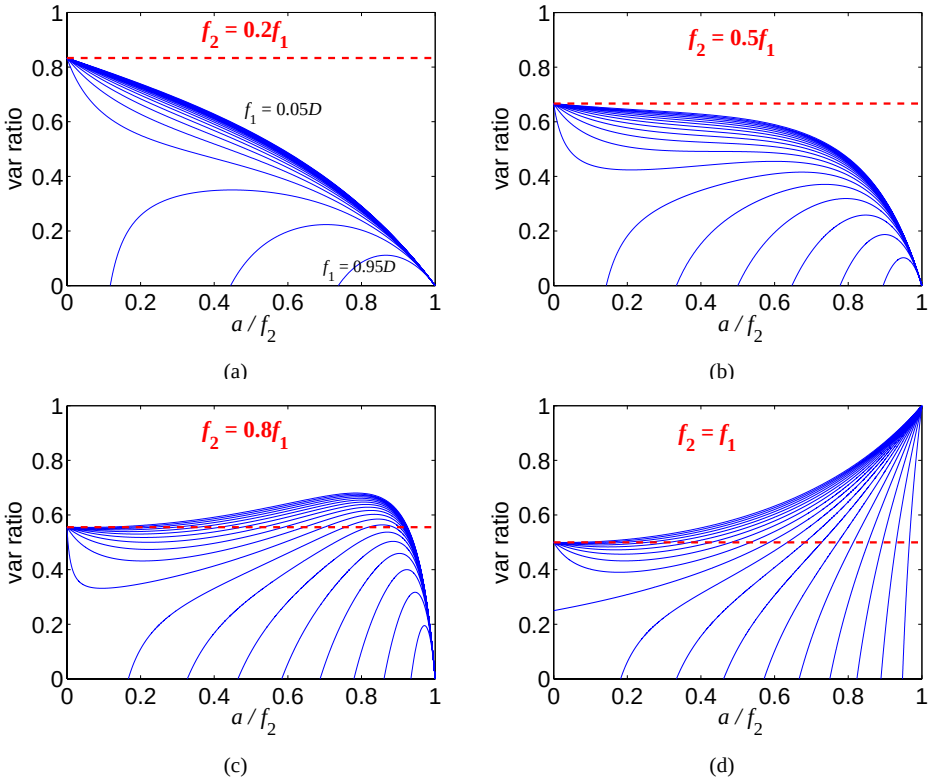


Figure 25
We plot V_B in (81) for the whole range of f_1, f_2, and a, assuming equal samples: $k_1 = k_2 = k$. (a), (b), (c) and (d) correspond to $f_2 = 0.2f_1$, $f_2 = 0.5f_1$, $f_2 = 0.8f_1$ and $f_2 = f_1$, respectively. Different curves are for different f_1's, ranging from $0.05D$ to $0.95D$ spaced at $0.05D$. The horizontal lines are $\frac{\max(f_1,f_2)}{f_1+f_2}$. Note that V_B in (81) is independent of D. We can see that for all cases, $V_B \le 1$ holds. $V_B = 1$ when $f_1 = f_2 = a$, a trivial case. When a/f_2 is small, $V_B \approx \frac{\max(f_1,f_2)}{f_1+f_2}$ holds well. It is also theoretically possible that V_B is zero when $d = D - f_1 - f_2 + a = 0$, or when $a = f_2 < f_1$.

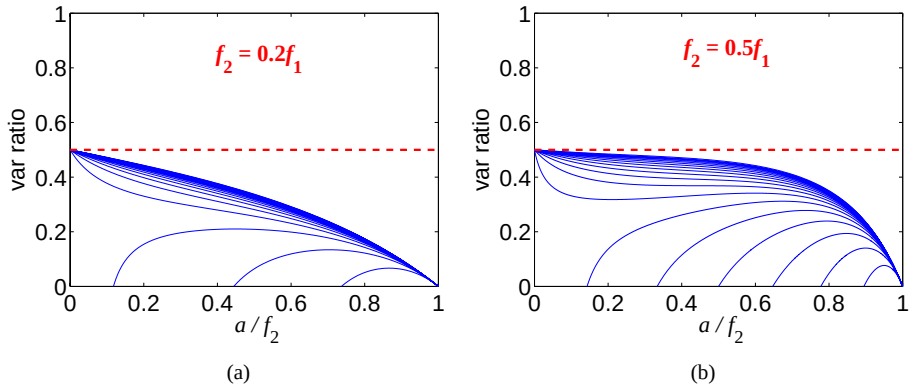


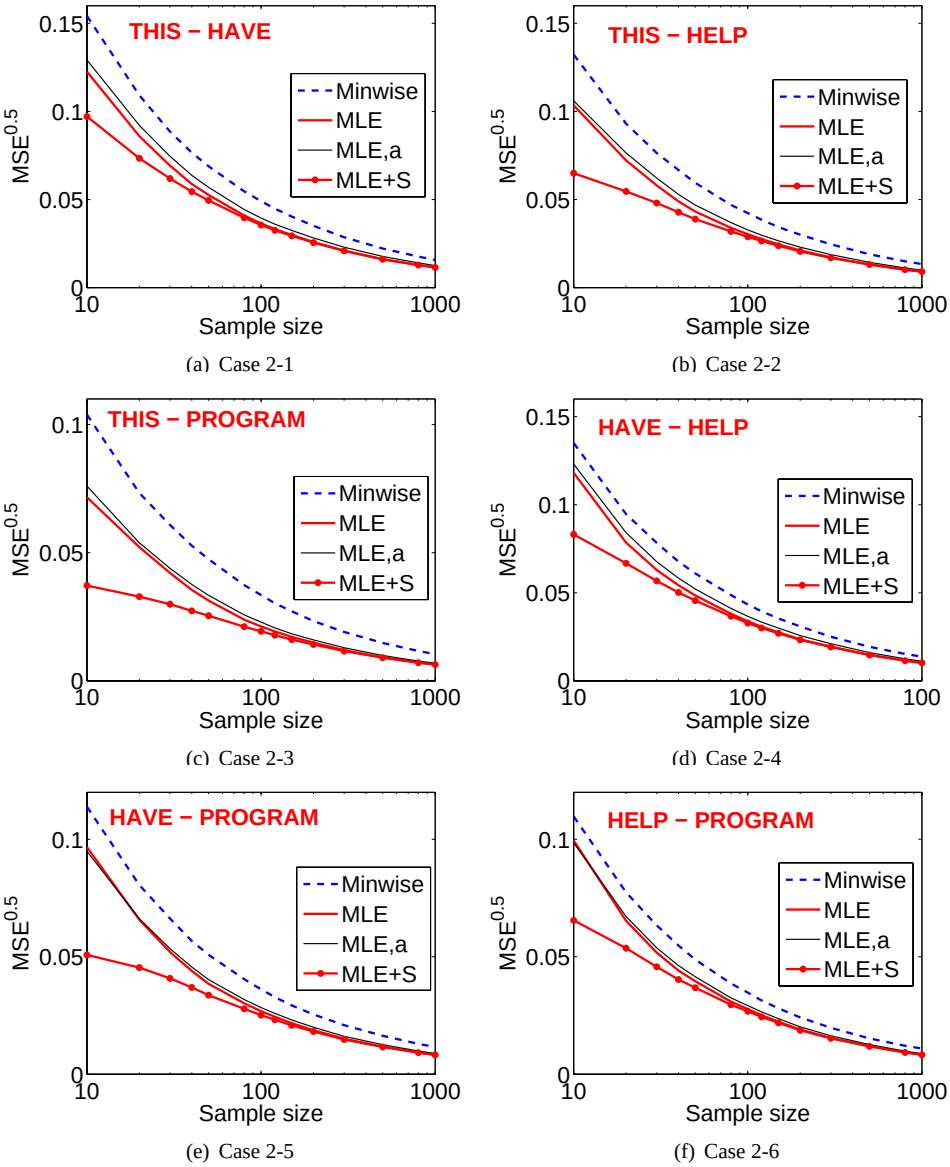
Figure 26
Compared with equal samples in Figure 25, proportional samples further reduce V_B.

which is equivalent to

$$\left(a^3(b-c)^2 + bc^2(b+c)^2 + a^2(2b+c)(b^2-bc+2c^2) + a(b+c)(b^3+4bc^2+c^2)\right) d$$
$$+ abc(b+c)(a+b+c)^2 \ge 0, \tag{87}$$

which always holds. The equality holds only when $b = c = 0$, i.e., $f_1 = f_2 = a$.

We can compare our estimated resemblance with Broder's sketch for the same small dataset in evaluating two-way associations, as given in Table 5. Figure 27 compares the MSE. Here we assume equal samples and later we will show that proportional samples could further improve the results. The figure shows that our algorithm is consistently better. The approximate MLE still gives very close answers to the exact MLE. Also, the simple "add-one" smoothing improves the estimations at low sampling rates, quite substantially.



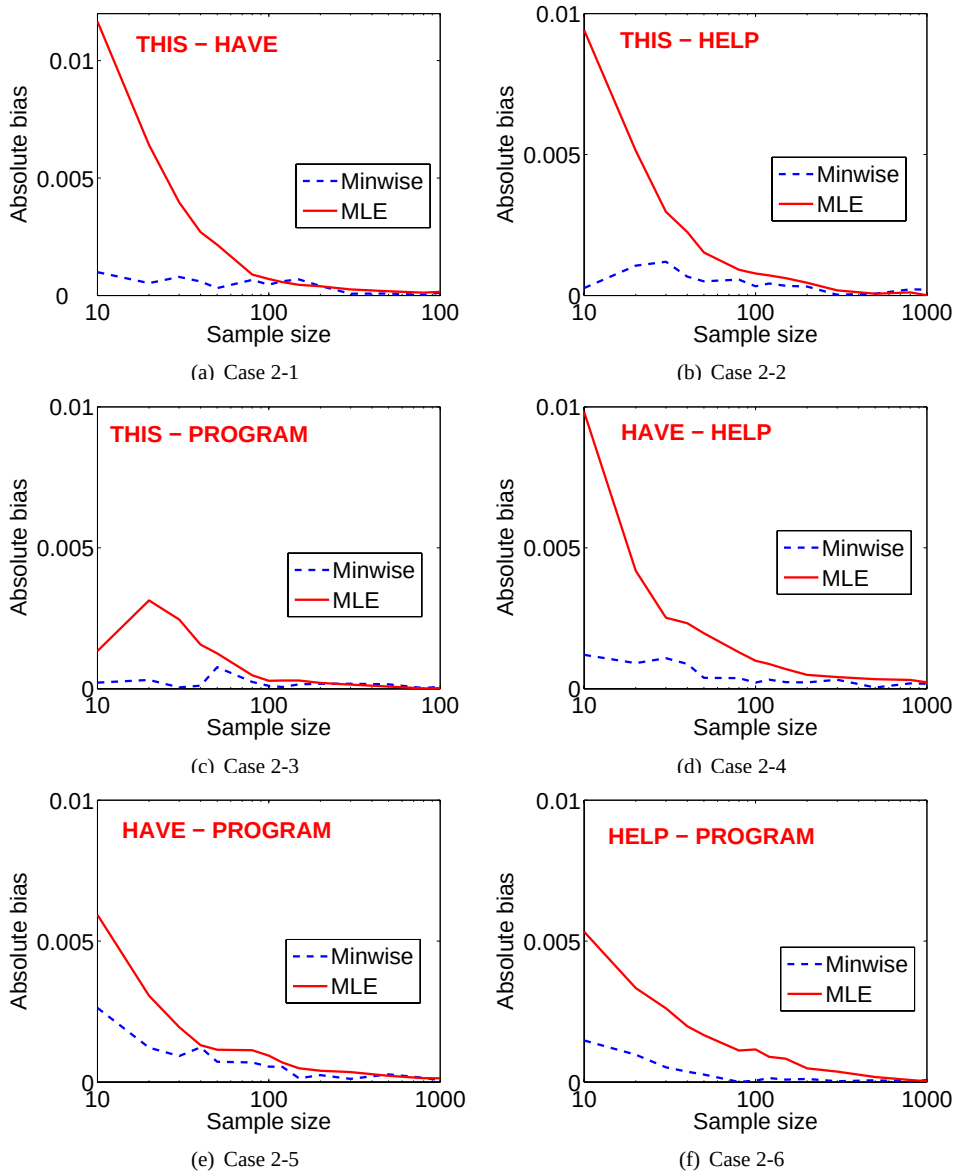
(a) Case 2-1 (b) Case 2-2

(c) Case 2-3 (d) Case 2-4

(e) Case 2-5 (f) Case 2-6

Figure 27
When estimating the resemblance, our algorithm gives consistently more accurate answers than Broder's sketch. In our experiments, Broder's "minwise" construction gives almost the same answers as the "sample-without-replacement" version, thus only the "minwise" results are presented here. The approximate MLE again gives very close answers to the exact MLE. Also, smoothing improves at low sampling rates.

Figure 28 illustrates the bias. As expected, our estimator for the resemblance has higher bias than Broder's sketch. However, since the absolute magnitude of the bias is very small compared with the MSE, also as expected, we can basically ignore the bias in our discussions.



(a) Case 2-1 (b) Case 2-2

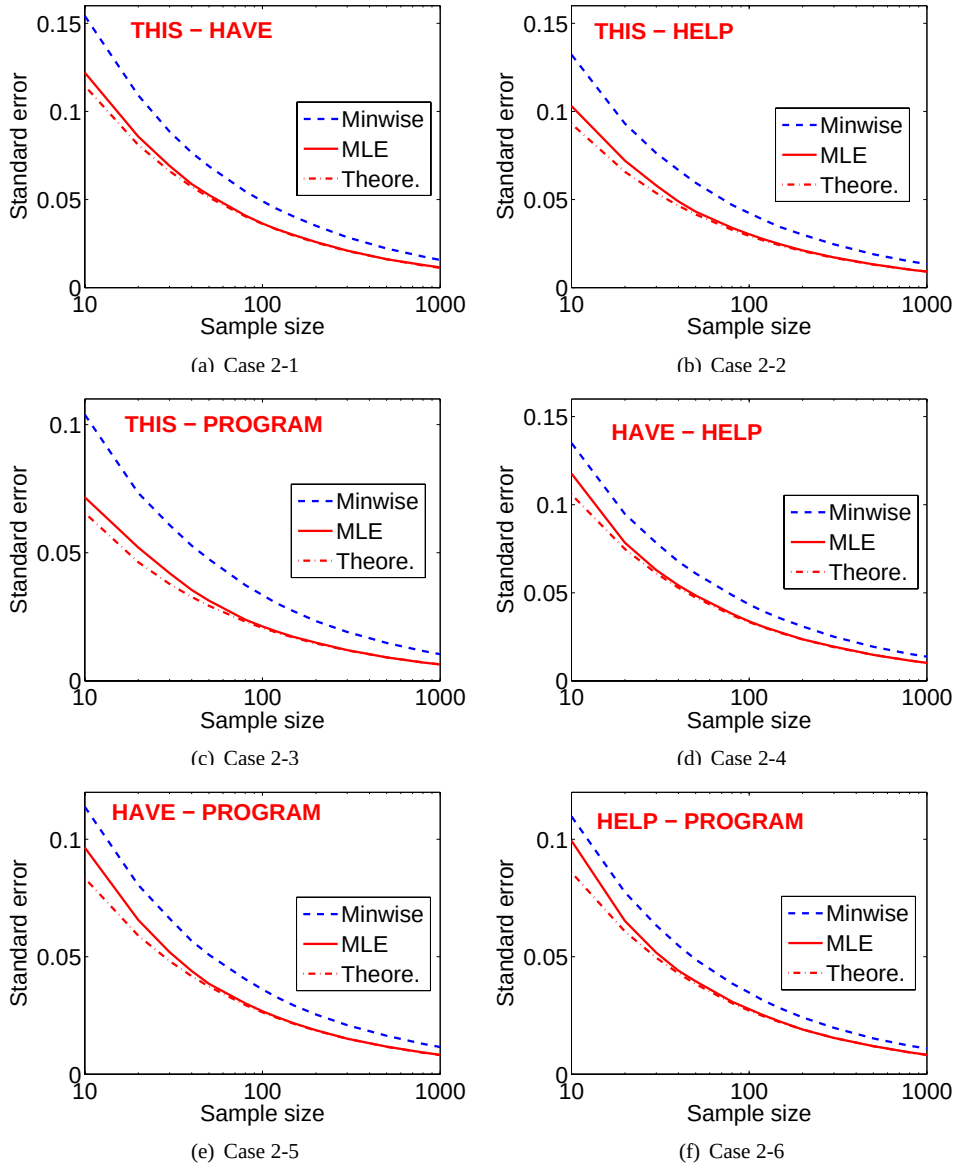
(c) Case 2-3 (d) Case 2-4

(e) Case 2-5 (f) Case 2-6

Figure 28
Our proposed MLE has higher bias than the "minwise" estimator because of the non-linearity of resemblance. However, the bias is very small compared with the MSE.

Figure 29 verifies that the variance of our estimator is always smaller than Broder's sketch. Our theoretical variance in (79) under-estimates the true variances for three reasons. First, The reciprocal of the Expected Fisher Information $\frac{1}{\mathrm{I}(a)}$ under-estimates the variance at very low sampling rates. Secondly, the approximation $\mathrm{E}\left(\frac{D}{D_s}\right) = \max\left(\frac{f_1}{k_1}, \frac{f_2}{k_2}\right)$ under-estimates the variance. Thirdly, the resemblance $R(a)$ is a convex function of a, hinting that the Delta Method also

under-estimates the variance. However, Figure 29 shows that the errors are not very big and become negligible after the sample size is large enough (e.g., 50). Therefore, we still regard our variance formula (79) reliable.



(a) Case 2-1 (b) Case 2-2

(c) Case 2-3 (d) Case 2-4

(e) Case 2-5 (f) Case 2-6

Figure 29
Our proposed estimator have consistently smaller variances than Broder's sketch. The theoretical variance, computed by (79) slightly under-estimates the true variance with small samples. Here we did not plot the theoretical variance for Broder's sketch because it is very close to the empirical curve.

Finally, in Figure 30, we show that with proportional samples, our algorithm further improves the estimates. In terms of the relative MSE, with equal samples, our estimators improves Broder's sketch by $30\% - 50\%$. With proportional samples, improvements become $40\% - 80\%$. The maximum possible improvement is 100%.

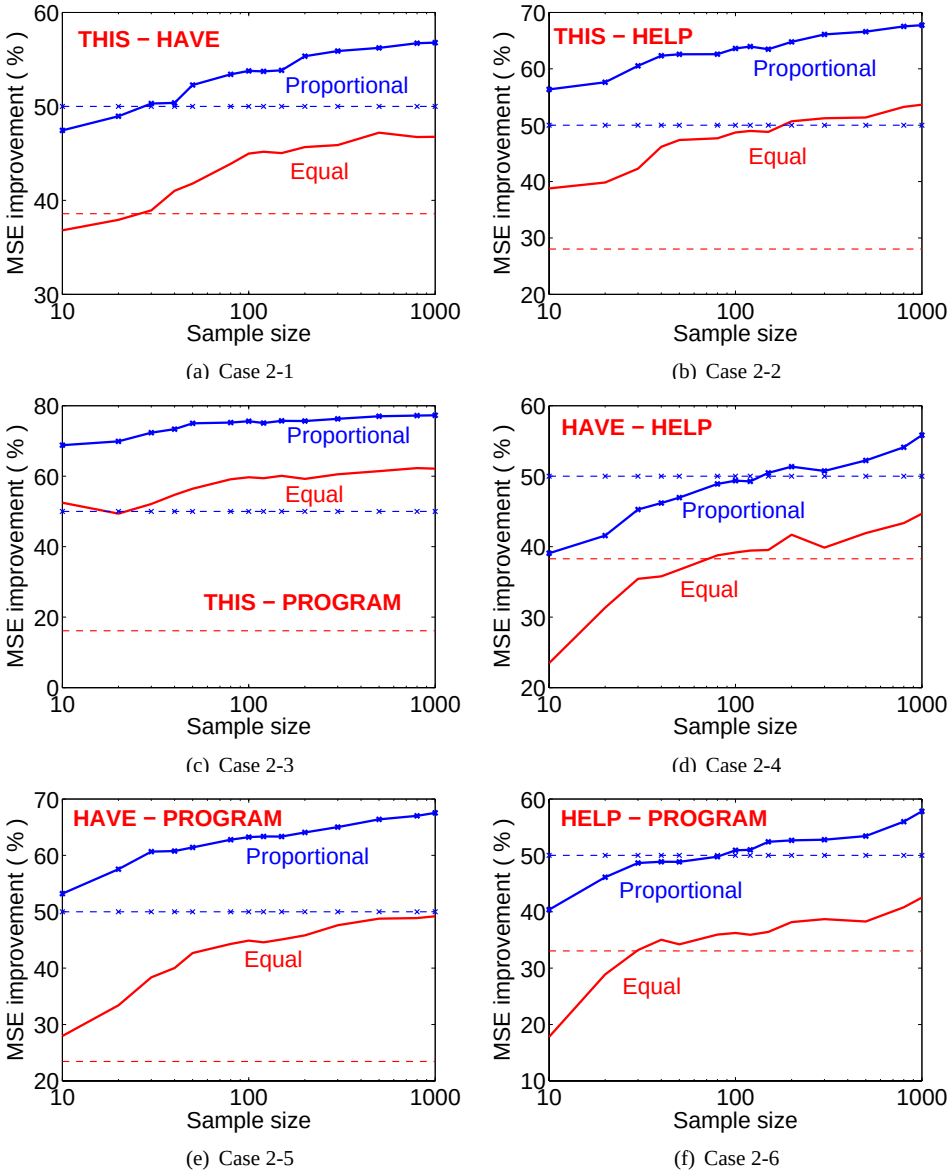


Figure 30

Compared with Broder's sketch, the relative MSE improvement should be, approximately, $\frac{\min(f_1,f_2)}{f_1+f_2}$ with equal samples, and $\frac{1}{2}$ with proportional samples. The two horizontal lines in each figure correspond to these two approximates. The actual improvements could be lower or higher. The figure verifies that proportional samples can considerably improve the accuracies.

9 Compare MLE with Random Projection in Estimating Cosine Angles

The random projection algorithm is also a very popular method for estimating vector similarity. Charikar (2002) treated the random projection algorithm as a special case for local sensitive hashing (LSH). Ravichandran et al. (2005) implemented it for estimating word associations.

For completeness, we repeat the basic theorem, first proved by Goemans and Williamson

(1995). Given two vectors v_1 and v_2 in D dimensions, and a random vector v_r whose entries consists of i.i.d. standard Normals, a hash function is defined as:

$$H_r(v_1) = \begin{cases} 1 & \text{if dot}(v_1, v_r) \geq 0 \\ 0 & \text{if dot}(v_1, v_r) < 0 \end{cases}. \tag{88}$$

Similary, we can define $H_r(v_2)$. Goemans and Williamson (1995) proved that

$$P(H_r(v_1) = H_r(v_2)) = 1 - \frac{\theta(v_1, v_2)}{\pi}, \tag{89}$$

where $\theta(v_1, v_2) = \cos^{-1}\left(\frac{\text{dot}(v_1,v_2)}{|v_1||v_2|}\right)$, is the angle between the two vectors v_1 and v_2.

Figure 31 gives an intuitive example why (89) is true in two dimensions.

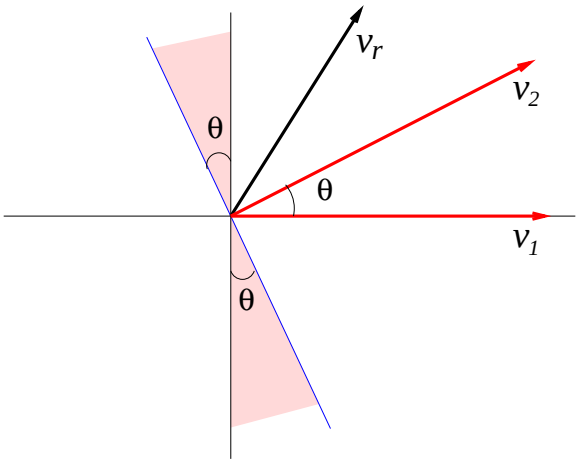


Figure 31
The angles between two vectors v_1 and v_2 is θ. Suppose there is a random vector v_r whose orientation is uniformly random in $[0, 2\pi]$. We can compute the hash $H_r(v_1)$ and $H_r(v_2)$ by (88). Suppose $\theta < \frac{\pi}{2}$ as shown in the figure. When v_r falls inside the shaded area, we know $H_r(v_1) \neq H_r(v_2)$, which occurs with a probability $\frac{2\theta}{2\pi}$, i.e., $P(H_r(v_1) = H_r(v_2)) = 1 - \frac{\theta(v_1,v_2)}{\pi}$, which holds even when $\theta > \frac{\pi}{2}$.

We can generate k such random vectors and estimate the probability $P(H_r(v_1) = H_r(v_2))$ as a binomial distribution, from which we can estimate the angle by

$$\hat{\theta}_{RP} = \left(1 - \hat{P}(H_r(v_1) = H_r(v_2))\right)\pi, \tag{90}$$

whose variance would be

$$\begin{aligned} \text{Var}\left(\hat{\theta}_{RP}\right) &= \frac{\pi^2}{k}\left(P(H_r(v_1) = H_r(v_2))\right)\left(1 - P(H_r(v_1) = H_r(v_2))\right) \\ &= \frac{\pi^2}{k}\left(1 - \frac{\theta}{\pi}\right)\left(\frac{\theta}{\pi}\right) = \frac{\theta(\pi - \theta)}{k}. \end{aligned} \tag{91}$$

Our sketch algorithm estimates a, from which we can compute the angle by

$$\hat{\theta}_{MLE} = \cos^{-1}\left(\frac{\hat{a}}{\sqrt{f_1 f_2}}\right), \tag{92}$$

whose variance is approximately

$$\text{Var}\left(\hat{\theta}_{MLE}\right) = \frac{\max\left(\frac{f_1}{k_1}, \frac{f_2}{k_2}\right)}{\frac{1}{a} + \frac{1}{f_1 - a} + \frac{1}{f_2 - a} + \frac{1}{D - f_1 - f_2 + a}} \frac{1}{f_1 f_2 - a^2}. \tag{93}$$

We can also define the ratio of $\text{Var}(\hat{\theta}_{MLE})$ over $\text{Var}(\hat{\theta}_{RP})$:

$$V_{RP} = \frac{\text{Var}(\hat{\theta}_{MLE})}{\text{Var}(\hat{\theta}_{RP})} = \frac{k \max\left(\frac{f_1}{k_1}, \frac{f_2}{k_2}\right)}{\frac{1}{a} + \frac{1}{f_1 - a} + \frac{1}{f_2 - a} + \frac{1}{D - f_1 - f_2 + a}} \frac{1}{f_1 f_2 - a^2} \frac{1}{\theta(\pi - \theta)}. \tag{94}$$

Figure 32 plots V_{RP} for $f_2 = 0.2 f_1$, $f_2 = 0.5 f_1$, $f_2 = 0.8 f_1$ and $f_2 = f_1$ for equal samples $k_1 = k_2 = k$. The plots show that the ratio V_{RP} is normally very small in most cases, except when $f_1 = f_2 = a$, which corresponds to $\theta = 0$, a trivial case. The figure implies that our estimator is better than random projection in estimating the angles. Proportional samples will further reduce the variance but we skip the plots here.

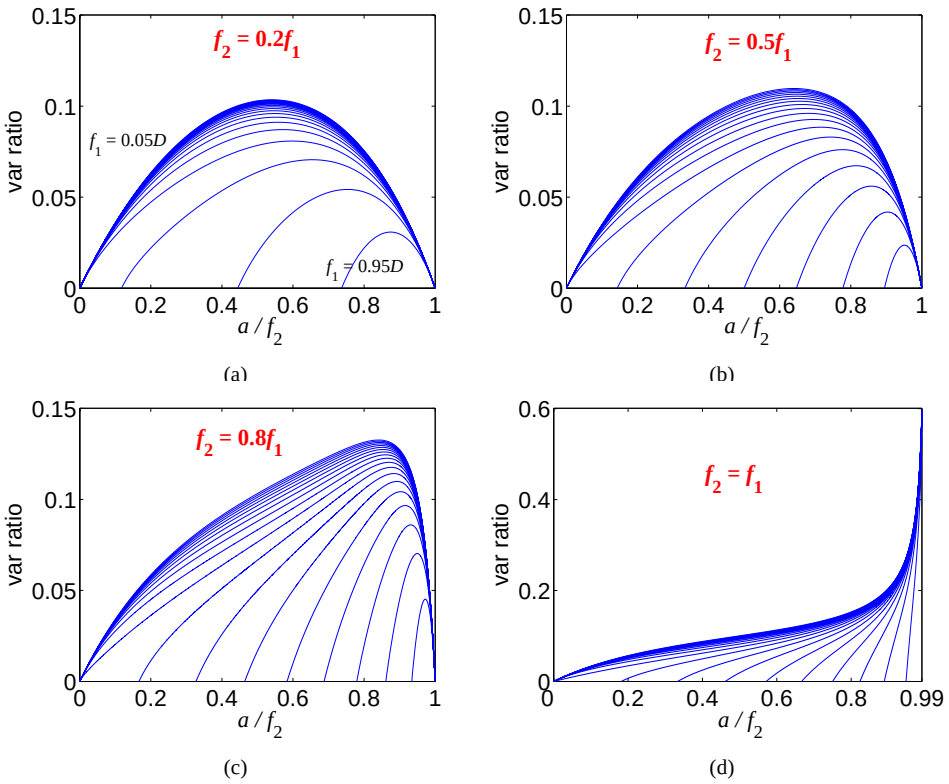
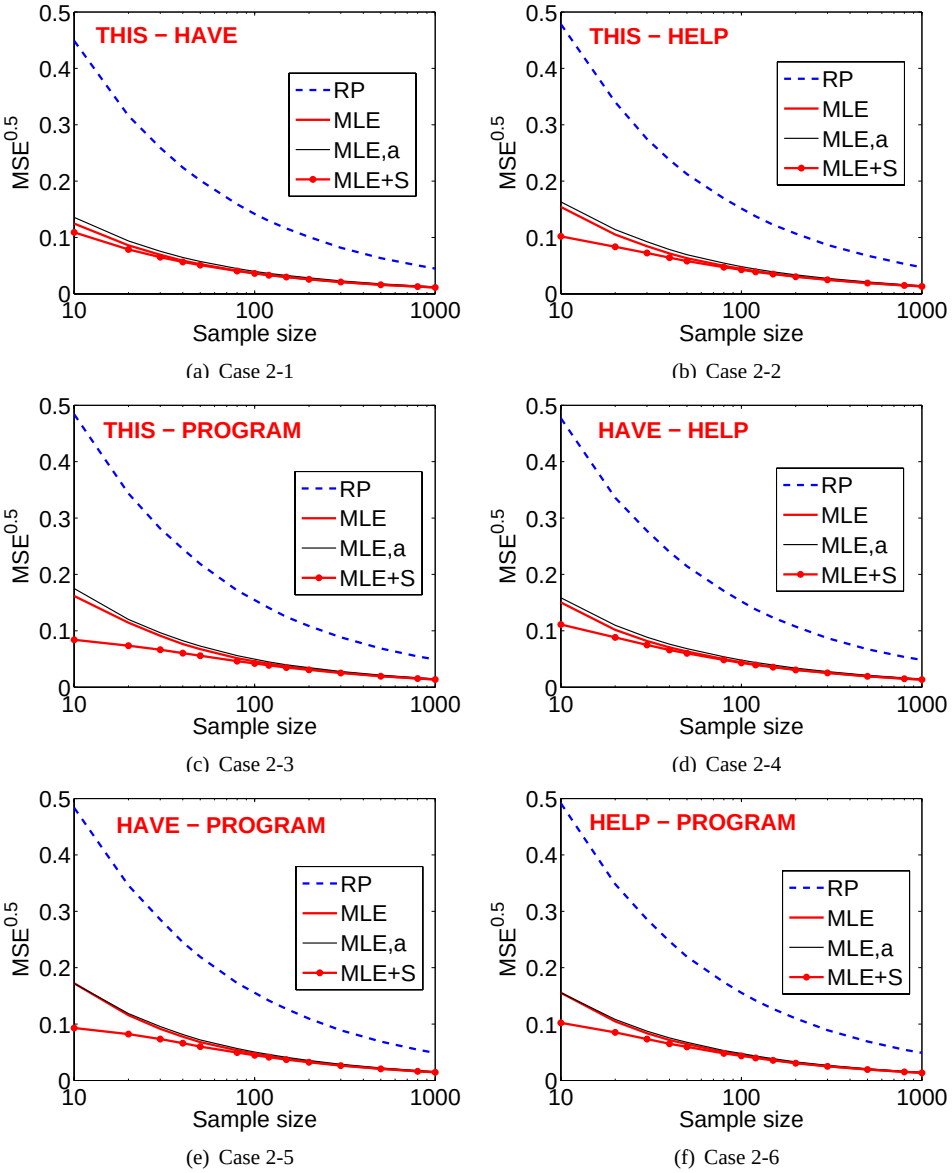


Figure 32
We plot V_{RP} in (94) for the whole range of f_1, f_2, and a, assuming equal samples: $k_1 = k_2 = k$. (a), (b), (c) and (d) correspond to $f_2 = 0.2 f_1$, $f_2 = 0.5 f_1$, $f_2 = 0.8 f_1$ and $f_2 = f_1$, respectively. Different curves are for different f_1's, ranging from $0.05D$ to $0.95D$ spaced at $0.05D$. The figure shows that the variance ratio $V_{RP} = \frac{\text{Var}(\hat{\theta}_{MLE})}{\text{Var}(\hat{\theta}_{RP})}$ is very small for all cases except when $f_1 = f_2 = a$, which is a singular case and the variance formula (93) derived by Delta Method is no longer applicable.

We compare the performance of our estimator with random projection in estimating the angles, using the same dataset of 4 words (6 word pairs) in Table 5. We first experiment with equal samples.

Figure 33 compares the MSE. Using our proposed MLE, we are able to estimate the angles much more accurately than random projection. The approximate MLE works almost as well as the exact MLE. Also, smoothing helps at low sampling rates.



(a) Case 2-1 (b) Case 2-2

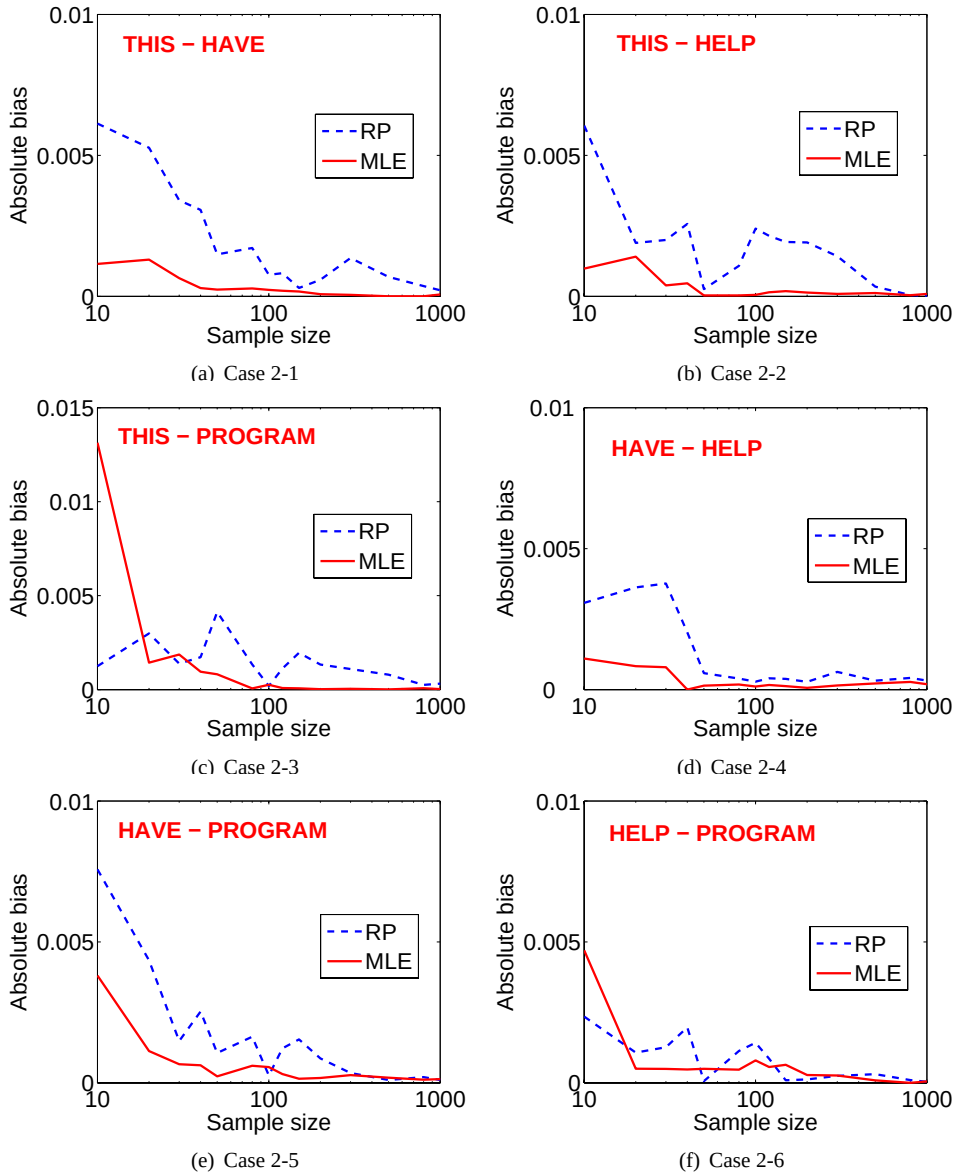
(c) Case 2-3 (d) Case 2-4

(e) Case 2-5 (f) Case 2-6

Figure 33
We compare random projection (RP) with our proposed MLE. Our method gives much more accurate estimates in terms of $\text{MSE}^{0.5}$. The approximate MLE works almost as well as the exact MLE. Also, smoothing helps at low sampling rates.

Figure 34 plots the absolute bias for both random projection and our proposed MLE. Both estimators are practically unbiased for these 6 cases.

Figure 35 compares the variance. Empirically, the proposed MLE has much lower variance than random projection, as expected. The theoretical variance for the proposed MLE fits the empirical variance quite well but there is no consistent trend whether the theoretical variance



(a) Case 2-1 (b) Case 2-2

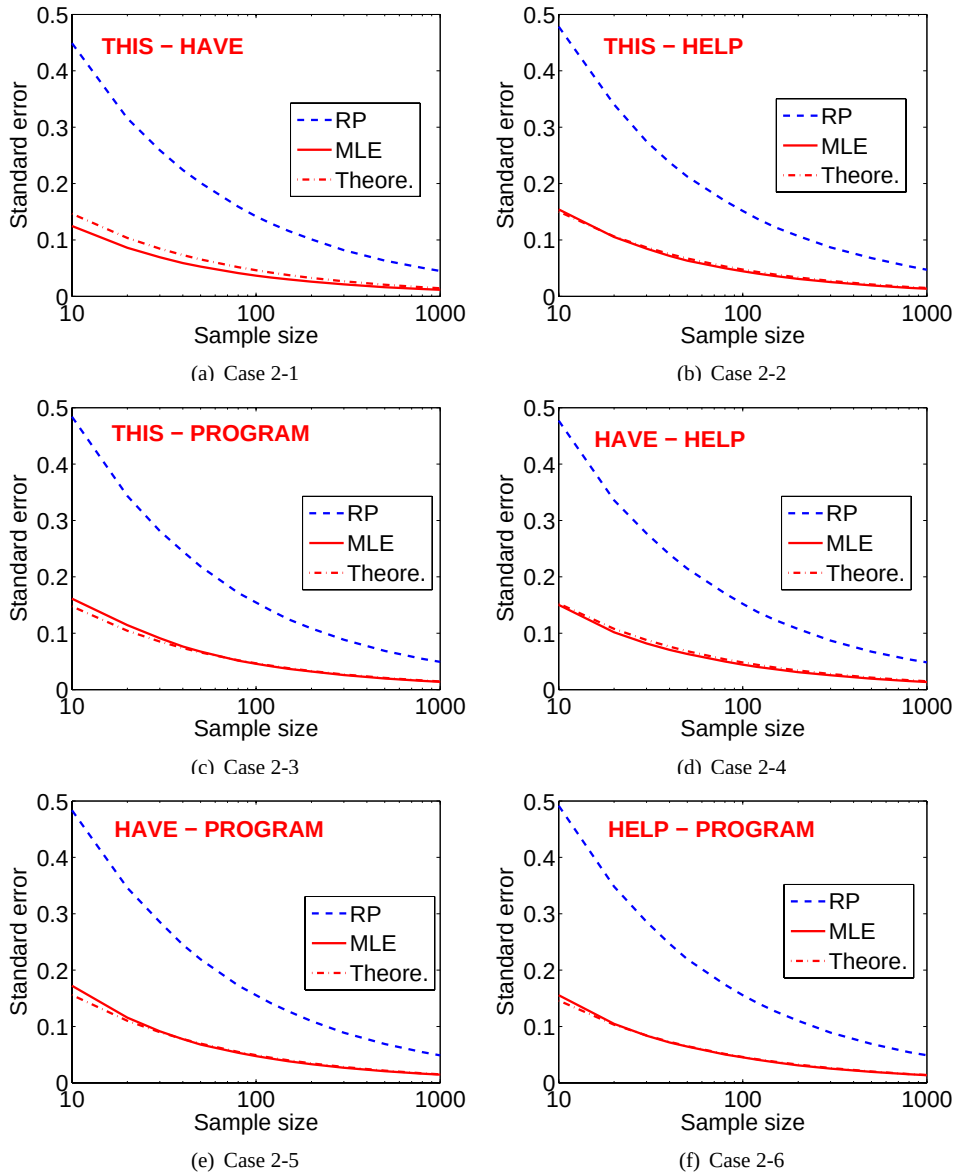
(c) Case 2-3 (d) Case 2-4

(e) Case 2-5 (f) Case 2-6

Figure 34
The biases in both random projection and our proposed MLE are very small compared with MSE.

under-estimate or over-estimate the true variance. Because the inverse cosine, $\cos^{-1}$, function is concave in $[0,\ \pi]$, the approximate variance (93) derived by Delta Method will in general over-estimate the true variance, while other factors (e.g., the approximation $\frac{D}{D_s} = \max\left(\frac{f_1}{k_1}, \frac{f_2}{k_2}\right)$) tend to under-estimate the variance.

Finally, we show in Figure 36 that with proportional samples, our estimators can further improve the accuracy, compared with random projection. Since the maximum possible relative improvement is 100% and even with equal samples our estimator is already much better than random projection, the additional improvement due to proportional samples may not appear to be very significant.



(a) Case 2-1 (b) Case 2-2

(c) Case 2-3 (d) Case 2-4

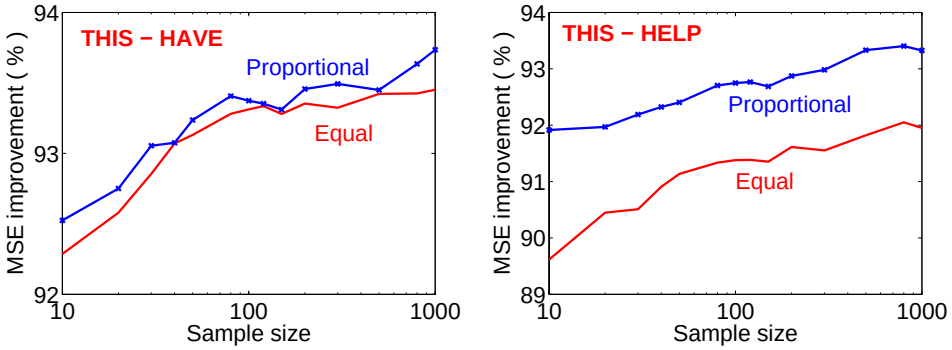
(e) Case 2-5 (f) Case 2-6

Figure 35
Empirically, the proposed MLE has much lower variance than random projection. The theoretical variance for the proposed MLE fits the empirical variance quite well.

10 Testing for Strong Two-way Associations

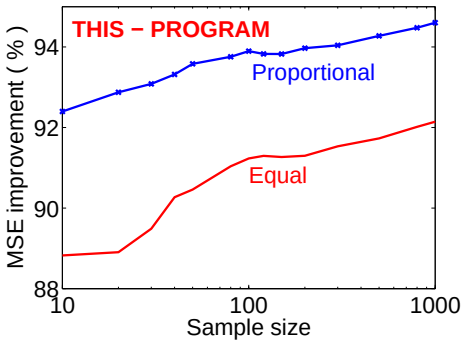
Statistical hypothesis testing has been widely used in NLP (Dunning, 1993; Moore, 2004) (Manning and Schutze, 1999, Section 5.3). We will first review four kinds of tests that are applicable for testing two-way associations: the Fisher's exact test, G^2 test, χ^2 test, and Poisson test.

We start with the Fisher's exact test. Assuming fixed margins f_1, f_2, the PMF of the co-

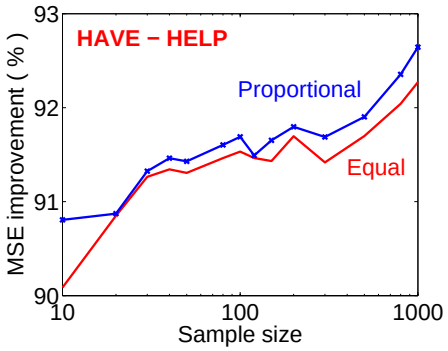


(a) Case 2-1

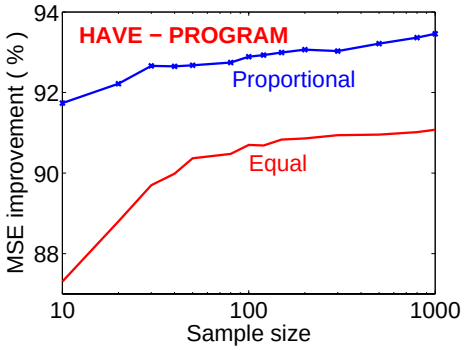
(b) Case 2-2



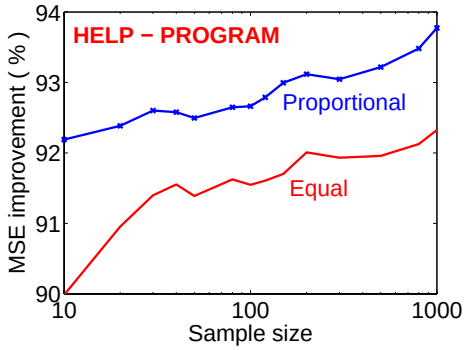
(c) Case 2-3



(d) Case 2-4



(e) Case 2-5



(f) Case 2-6

Figure 36
With proportional samples, the proposed MLE can further improvement the MSE compared with random projection. Note that the maximum possible relative improvement is 100%.

occurrences of W_1 and W_2 is

$$P(a=t;\text{Fisher}) = \frac{\binom{f_1}{t}\binom{D-f_1}{f_2-t}}{\binom{D}{f_2}} = \prod_{i=0}^{t-1}\frac{(f_1-i)(f_2-i)}{(t-i)(D-f_1+t-i)}\prod_{i=0}^{f_1-t-1}\frac{D-f_2-i}{D-i}$$
$$= \exp\left(\sum_{i=0}^{t-1}\log\frac{(f_1-i)(f_2-i)}{(t-i)(D-f_1+t-i)} + \sum_{i=0}^{f_1-t-1}\log\frac{D-f_2-i}{D-i}\right). \quad (95)$$

In (95), we expand the PMF into a product form and a log-summation-exp form for nu-

merical reasons because factorial can easily exceed the largest machine number (e.g., $100! = 9.33 \times 10^{157}$). In most cases, the product form in (95) suffices but occasionally we will need to use the more expensive log-summation-exp form to avoid overflow. There are other more efficient (but approximate) alternatives to compute the PMF such as the Binet's log gamma formulas, (Weisstein, 2005a, Web resource), which was also used in (Moore, 2004).

To test whether W_1 and W_2 are strongly correlated, we can compute the "p-value,"

$$p\text{-value} = \sum_{t=a}^{\min(f_1,f_2)} P(a = t; \text{Fisher}). \tag{96}$$

If p-value $< \alpha$, the significance level, we accept the alternative hypothesis that W_1 and W_2 are strongly correlated. Since we are mostly interested in positive correlations, we will focus our discussion on one-sided tests, i.e., only testing whether two words are strongly positively-correlated.

Alternatively, we can compute a "critical value," a_{cr}, based on the level of significance,

$$P(a \geq a_{cr}) = \sum_{t=a_{cr}}^{\min(f_1,f_2)} P(a = t; \text{Fisher}) = \alpha \tag{97}$$

If the observed value a exceed a_{cr}, we accept that W_1 and W_2 are significant.

The Fisher's exact test is usually considered be to "exact" and most suitable when the D is small. However, we should also be aware of its limitations:

1. The "exactness" is based on assuming that the margins, f_1 and f_2, are fixed. In other words, the Fisher's exact test is a "conditional test." When we try to estimate a, we take the advantage of knowing the margins. However, in hypothesis testing, the fixed-margins assumption may not be always appropriate because at the corpus level, the margins are also random. The argument about the "conditional test" v.s. "unconditional test" is still a debatable issue (Agresti, 2002, Chapter 3.5.5 3.5.6).

2. The PMF in the Fisher's exact test, i.e., the hypergeometric distribution, is highly discrete, which makes it not possible to achieve the exact significance level and leads to conservative tests. There are some procedures to (partially) overcome these shortcomings, such as the "randomization extension" or the "mid-p-value adjustment" (Agresti, 2002, Chapter 3.5.4). However, the problem of discreteness is not that significant for multiple (simultaneous) hypothesis testing especially when the simplest multiple testing method, the Bonferroni's method, which leads to very small significance levels, is used. We will discuss more about multiple testing later.

3. The Fisher's exact test is computationally very expensive. Although there are accurate log-Gamma approximations that can speed up in computing the PMF, one still has to compute the cumulative probability as in (96) and (97). Normally one can write the PMF $P(a = t - 1)$ recursively in terms of $P(a = t)$ to save computations, but we need to be very careful of error propagations because $P(a = \min(f_1, f_2))$ is normally nearly zero.

We will assume that we do not worry about the "conditional v.s. unconditional" and "discreteness" issues, so that we can use the Fisher's exact as the "gold standard" for other hypothesis testing methods.

Next, we will review the G^2 test and χ^2 test. We have already seen the LLR statistics, which

is $\frac{1}{2}G^2$, i.e.,

$$G^2 = G^2(a) = 2a \log \frac{Da}{f_1 f_2} + 2(f_1 - a) \log \frac{(f_1 - a)D}{(D - f_2)f_1} + 2(f_2 - a) \log \frac{(f_2 - a)D}{(D - f_1)f_2} + 2(D - f_1 - f_2 + a) \log \frac{(D - f_1 - f_2 + a)D}{(D - f_1)(D - f_2)}. \quad (98)$$

The χ^2 statistics can be written as

$$\chi^2 = \chi^2(a) = \frac{\left(a - \frac{f_1 f_2}{D}\right)^2}{\frac{f_1 f_2}{D}} + \frac{\left(f_1 - a - \frac{f_2(D-f_1)}{D}\right)^2}{\frac{f_2(D-f_1)}{D}} + \frac{\left(f_2 - a - \frac{f_1(D-f_2)}{D}\right)^2}{\frac{f_1(D-f_2)}{D}} + \frac{\left(D - f_1 - f_2 + a - \frac{(D-f_1)(D-f_2)}{D}\right)^2}{\frac{(D-f_1)(D-f_2)}{D}}. \quad (99)$$

Here we have implicitly used the "fixed-margins" assumption when we write G^2 (or χ^2) as a function of a only.

It is well-known that both G^2 and χ^2 converge to the Chi-squared distribution (which is also represented as χ^2). In fact χ^2 is basically the second-order Taylor expansion of G^2 (Dunning, 1993)(Shao, 1999, Chapter 6.4)(Agresti, 2002, Chapter 14.3.3). Therefore, it should be almost always the case that G^2 is more accurate and converges faster than χ^2, especially at very small significance level because G^2 has heavier tail than χ^2.

Note that the only condition required for convergence is that $D \to \infty$, which is very well approximated in NLP applications.

We can similarly compute the critical values, a_{cr} for both G^2 and χ^2 tests by solving

$$G^2(a_{cr;G^2}) - \chi^2_{1,1-\alpha} = 0, \quad (100)$$

$$\chi^2(a_{cr;\chi^2}) - \chi^2_{1,1-\alpha} = 0. \quad (101)$$

where $\chi^2_{1,1-\alpha}$ is the Chi-squared critical value at α significance level and one degree of freedom.

We have to solve for $a_{cr;G^2}$ numerically. However, there is an exact solution for $a_{cr;\chi^2}$:

$$a_{cr;\chi^2} = \left(\frac{f_1 f_2}{D}\right) + \sqrt{\chi^2_{1,1-\alpha}\left(\frac{f_1 f_2}{D}\right)\frac{(D - f_1)(D - f_2)}{D^2}}, \quad (102)$$

which can be used as the initial value for $a_{cr;G^2}$.

Finally, we introduce the Poisson test, which is not as widely-used but it has certain advantages. The concept of Poisson approximation is well-known. For example, a binomial with probability p and counts n can be often well-approximated by Poisson with parameter $\lambda = np$ if n is large and p is small. For more details about Poisson approximations, see (Ross, 1996, Chapter 2, 10).

For the words we are interested, it is often the case that the document frequencies are much smaller than D, the corpus size. Therefore, under the independence assumption, the distribution of a can be approximated as a Poisson with parameter $\lambda = \frac{f_1 f_2}{D}$ (compared with the more familiar Poisson approximation for binomial, $\frac{\min(f_1, f_2)}{D}$ can be considered as "p," and $\max(f_1, f_2)$ as "n"). The PMF of a according to a Poisson distribution would be

$$P(a = t; \text{Poisson}) = \frac{\exp\left(-\frac{f_1 f_2}{D}\right)\left(\frac{f_1 f_2}{D}\right)^t}{t!}, \quad (103)$$

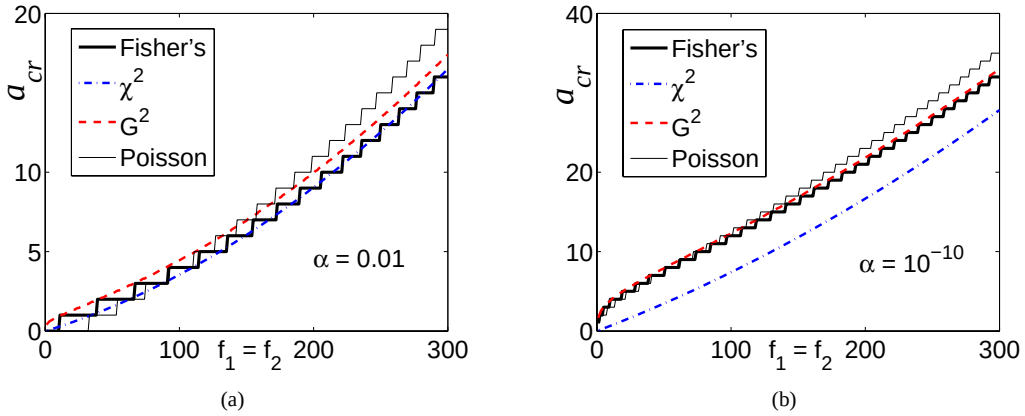


Figure 37
We compare the critical value, a_{cr}, for four different kinds of tests. $D = 10^5$, $f_1 = f_2$, and two significance levels, $\alpha = 0.01$ (a) and $\alpha = 10^{-10}$ (b). The figures illustrate that at very small significance levels, for relatively infrequent words, the Fisher's exact test and the G^2 test are very similar. The Poisson test matches the Fisher's test fairly well at low document frequencies. The χ^2 test, however, differ significantly from all other three tests.

which is much simpler than the PMF for the Fisher's exact test.

We compared the a_{cr} for all four tests for $D = 10^5$, $f_1 = f_2 = 1$ to 300, at two significance levels, $\alpha = 0.01$ and $\alpha = 10^{-10}$. The results are plotted in Figure 37.

One may ask, isn't $\alpha = 10^{-10}$ too low as a significance level? Not necessarily. When we simultaneously test many pairs of words, using a significance level = 0.01 may result in falsely accepting too many pairs to be strongly associated. Here we introduce the concept of "family-wise error rate" (FWER), which is the rate at which a statistical test would be expected to produce one or more false positives among a class (family) of tests, under the null hypothesis. The Bonferroni's method (Shao, 1999, Chapter 7.5.1) is a very conservative approach that achieve the designed FWER by dividing the level of significance by the total number of tests, i.e., $\frac{\alpha}{N}$. In our understanding, N does not necessarily have to the number of tests in the current experiment. Instead, it should reflect that in total how many tests one intend to perform.

A relatively new approach of multiple testing is the false discovery rate (FDR) method (Benjamini and Hochberg, 1995). FDR controls the expected proportion of false positives (false discoveries) and is in general much more powerful than the Bonferroni's method. Note that the FDR = FWER if all null hypotheses are true. The FDR method has become very popular recently, e.g., (Efron and Tibshirani, 2002; Storey, 2003).

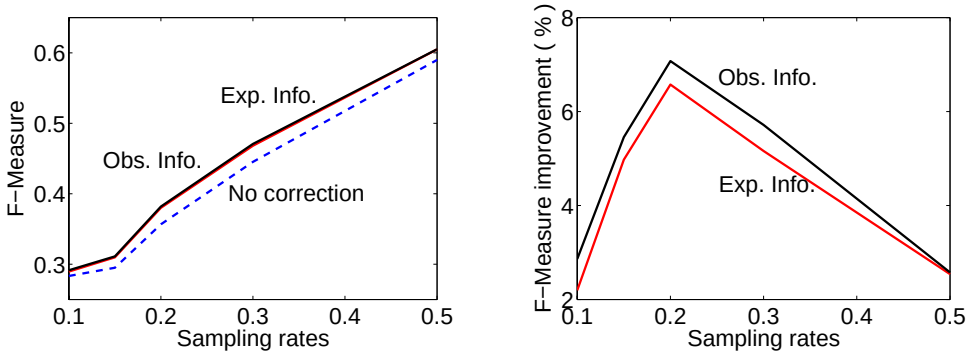
We decide to perform multiple G^2 tests on a set of relatively less frequent words, using the Bonferroni's method. Because we have to use the estimated a to compute the G^2 statistics, we would like to know the effects of sampling on the hypotheses testing results. Directly related to sequential sampling, there is also the issue of sequential hypothesis testing (Siegmund, 1985, Chapter III, IV), which we do not delve in.

We now describe our experiment on multiple tests. Because we are mainly interested in relatively rare words, in our experiment, we select those words whose document frequencies are ≤ 200 from the dataset, resulting in 1953 word pairs. At a significance level of 0.01 (with Bonferrorni's correction, the actual significance is $\frac{0.01}{1953} = 5.12 \times 10^{-6}$), the G^2 tests report that 305 pairs are considered as strongly correlated. Next we performed the multiple G^2 testing on the same 1953 word pairs using the estimated a values. We report precision and recall at different sampling rates. Since we did not vary the significance levels, there is no obvious trade-

off between the precision curves and the recall curves. To combine the precision and recall into a single measure, we report the "F-measures" (Manning and Schutze, 1999, Chapter 8.1). A common choice of F-measure is that F-measure = $2\frac{\text{precision}\times\text{recall}}{\text{precision}+\text{recall}}$.

We report the F-measures v.s. sampling rates in Figure **??**. Because G^2 (LLR) is highly non-linear, we correct the estimated G^2 by the "Gamma correction." The variances are computed by both the Expected Fisher Information and Observed Fisher Information.

Figure **??** (a) shows the "Gamma correction" can improve the F-measure. Figure **??**(b) plots the relative improvement using "Gamma correction" with respect to the uncorrected G^2 test. We can see that with "Gamma correction," we can improve the G^2 accuracy by up to 7%. In particular, using the variance computed by the Observed Fisher Information performs slightly better than the variance computed by the Expected Fisher Information for this dataset.



(a) (b)

Figure 38
(a): Three F-measure curves with respect to the sampling rates indicate that correcting the G^2 statistics (by "Gamma correction") can improve the hypothesis testing accuracy. (b): With respect to the non-corrected G^2 test, The Gamma correction can improve up to 7%. Using the Observed Fisher Information gives slightly better results than using the Expected Fisher Information.

11 Extension to Multi-way Associations

Many applications involve multi-way associations as opposed to two-way associations. For example, in Databases and Web search, user queries are not necessarily limited to two "words." The "Governator" example in Table 3, for example, made use of three-way associations in addition to two-way associations. Fortunately, our sketch construction and estimation algorithm can be naturally extended to multi-way associations.

In this section, we will show that estimating multi-way associations using our sketch algorithm amounts to a convex optimization problem for the exact MLE. We will present an algorithm to analyze the estimation variance. We will also compare our proposed MLE estimator with two baselines.

11.1 Multi-way Sketches

We will need some more notation to discuss multi-way assocations. Suppose we are interested in the associations among m words, denoted by W_1, W_2, ..., W_m. The document frequencies of these m words are f_1, f_2, ..., and f_m, which are also the lengths of the postings P_1, P_2, ..., P_m.

There are $N = 2^m$ combinations of associations, denoted by $x_1, x_2, ..., x_N$. For example,

we can define

$$
\begin{aligned}
x_1 &= |\mathrm{P}_1 \cap \mathrm{P}_2 \cap ... \cap \mathrm{P}_{m-1} \cap \mathrm{P}_m|, \\
x_2 &= |\mathrm{P}_1 \cap \mathrm{P}_2 \cap ... \cap \mathrm{P}_{m-1} \cap \neg \mathrm{P}_m|, \\
x_3 &= |\mathrm{P}_1 \cap \mathrm{P}_2 \cap ... \cap \neg \mathrm{P}_{m-1} \cap \mathrm{P}_m|, \\
x_4 &= |\mathrm{P}_1 \cap \mathrm{P}_2 \cap ... \cap \neg \mathrm{P}_{m-1} \cap \neg \mathrm{P}_m|, \\
&... \\
x_{N-1} &= |\neg \mathrm{P}_1 \cap \neg \mathrm{P}_2 \cap ... \cap \neg \mathrm{P}_{m-1} \cap \mathrm{P}_m|, \\
x_N &= |\neg \mathrm{P}_1 \cap \neg \mathrm{P}_2 \cap ... \cap \neg \mathrm{P}_{m-1} \cap \neg \mathrm{P}_m|,
\end{aligned} \tag{104}
$$

which can be directly related to the binary representation of integers. Table 7 gives two examples for $m = 2$ and $m = 3$, respectively.

Table 7
We number the associations, $x_1, x_2, ..., x_N$, using binary numbers. For each word W_i, a "0" indicates that documents containing W_i are in the intersections, an "1" indicates the complement (i.e., the documents not containing word W_i) are in the intersection. This way, the binary representation of the subscript of x_i minus 1 (i.e, $i - 1$) corresponds to the set intersections. For example, when $m = 3$, the binary representation of 3 is "0 1 1," indicating $x_4 = |\mathrm{P}_1 \cap \neg \mathrm{P}_2 \cap \neg \mathrm{P}_3|$.

(a) $m = 2$

	W_1	W_2
x_1	0	0
x_2	0	1
x_3	1	0
x_4	1	1

(b) $m = 3$

	W_1	W_2	W_3
x_1	0	0	0
x_2	0	0	1
x_3	0	1	0
x_4	0	1	1
x_5	1	0	0
x_6	1	0	1
x_7	1	1	0
x_8	1	1	1

For convenience, we introduce some vector and matrix notation. We denote $\mathbf{X} = [x_1, x_2, ..., x_N]^{\mathrm{T}}$. $\mathbf{F} = [f_1, f_2, ..., f_m, D]^{\mathrm{T}}$ is a vector of document frequencies (margins) and the total number of documents. Suppose all the margins and the corpus size are known, we can write down the constraints in terms of a linear matrix equation as

$$\mathbf{A}\mathbf{X} = \mathbf{F}, \tag{105}$$

where $\mathbf{A}$ is the constraint matrix. If necessary, we can use $\mathbf{A}^{(m)}$ to identify $\mathbf{A}$ for different m values. For example, when $m = 2$ or $m = 3$, this matrix becomes

$$
\mathbf{A}^{(2)} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \qquad \mathbf{A}^{(3)} = \begin{bmatrix} 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix}. \tag{106}
$$

The sampling procedure for multi-way associations is very similar to that for two-way associations. For each word W_i, we sample the first k_i elements from its sorted postings, P_i, to form a sketch, K_i. We can compute

$$D_s = \min\{\mathrm{K}_{1(k_1)}, \mathrm{K}_{2(k_2)}, ..., \mathrm{K}_{m(k_m)}\}, \tag{107}$$

where $\mathrm{K}_{i(k_i)}$ denotes the last element in the sketch K_i. After removing the elements in all K_i's that are larger than D_s, we can then intersect these m trimmed sketches to generate the sample table counts. The samples are denoted as $\mathbf{S} = [s_1, s_2, ..., s_N]^{\mathrm{T}}$.

Conditional on D_s, the samples $\mathbf{S}$ are statistically equivalent to D_s random samples over documents from the corpus. The corresponding conditional PMF would be

$$P(\mathbf{S}|D_s; \mathbf{X}) = \frac{\binom{x_1}{s_1}\binom{x_2}{s_2}...\binom{x_N}{s_N}}{\binom{D}{Ds}} \propto \prod_{i=1}^{N}\prod_{j=0}^{s_i-1}(x_i - j), \tag{108}$$

and the $\log$ likelihood would be

$$\log P(\mathbf{S}|D_s; \mathbf{X}) \propto Q = \sum_{i=1}^{N}\sum_{j=0}^{s_i-1}\log(x_i - j), \tag{109}$$

which is a concave function.

A partial likelihood MLE solution, i.e., the $\hat{\mathbf{X}}$ that maximizes $\log P(\mathbf{S}|D_s; \hat{\mathbf{X}})$, will again be adopted, which leads to a convex optimization problem. But first, we shall discuss two baseline estimators.

11.2 Baseline Independence Estimator

Assuming independence, an estimator of x_1 would be

$$\hat{x}_{1,IND} = D\prod_{i=1}^{m}\frac{f_i}{D}, \tag{110}$$

which can be easily proved using a conditional expectation argument.

We have seen that, according to a hypergeometric distribution,

$$\mathrm{E}(|\mathrm{P}_i \cap \mathrm{P}_j|) = \frac{f_i f_j}{D}. \tag{111}$$

Therefore,

$$\begin{aligned}
\mathrm{E}(x_1) &= \mathrm{E}(|\mathrm{P}_1 \cap \mathrm{P}_2| \cap ... \cap \mathrm{P}_m|) = \mathrm{E}(|\cap_{i=1}^{m}\mathrm{P}_i|)\\
&= \mathrm{E}(\mathrm{E}(|\mathrm{P}_1 \cap (\cap_{i=2}^{m}\mathrm{P}_i)\,||\,(\cap_{i=2}^{m}\mathrm{P}_i)))\\
&= \frac{f_1}{D}\mathrm{E}(|\cap_{i=2}^{m}\mathrm{P}_i|)\\
&...\\
&= \frac{f_1 f_2 ... f_{m-2}}{D^{m-2}}\mathrm{E}(|\mathrm{P}_{m-1} \cap \mathrm{P}_m|)\\
&= D\prod_{i=1}^{m}\frac{f_i}{D}.
\end{aligned} \tag{112}$$

We can estimate x_2 etc., in a similar fashion. In fact, we can ignore the expectation operation and simply treat $|\mathrm{P}_i \cap \mathrm{P}_j| = \frac{f_i f_j}{D}$. This way, we can write down the expressions for $\hat{x}_2$ etc. very easily using set operations and the results will still be correct. For example,

$$\begin{aligned}
\hat{x}_{4,IND} &= \mathrm{E}(|\mathrm{P}_1 \cap \mathrm{P}_2 \cap ... \cap \neg\mathrm{P}_{m-1} \cap \neg\mathrm{P}_m|)\\
&\overset{\text{treat as}}{=\!=} |(\mathrm{P}_1 \cap \mathrm{P}_2 \cap ... \cap \mathrm{P}_{m-2}) \cap (\neg\mathrm{P}_{m-1} \cap \neg\mathrm{P}_m)|\\
&= \left(D\prod_{i=1}^{m-2}\frac{f_i}{D}\right)\left(D - f_{m-1} - f_m + \frac{f_{m-1}f_m}{D}\right)\frac{1}{D}.
\end{aligned} \tag{113}$$

11.3 Baseline Margin-free Estimator

Ignoring the margin constraints, the conditional PMF $P(\mathbf{S}|D_s;\mathbf{X})$ becomes the multivariate hypergeometric distribution, based on which we can derive the margin-free estimators to be

$$\mathrm{E}(s_i) = \frac{D_s}{D}x_i, \tag{114}$$

$$\hat{x}_{i,MF} = \frac{D}{D_s}s_i, \tag{115}$$

$$\mathrm{Var}(\hat{x}_{i,MF}) = \frac{D}{D_s}\frac{1}{\frac{1}{x_i}+\frac{1}{D-x_i}}\frac{D-Ds}{D-1} \tag{116}$$

We can see that the margin-free estimator remains its simplicity in the multi-way case. When we work with two-way associations, the resultant exact MLE is the solution to a cubic equation; not a big deal. In addition, the approximate MLE for the two-way association is a solution to a quadratic equation, which is very close to the exact MLE.

Next, we will show how to solve for the exact MLE, considering the margin constraints.

11.4 The Exact MLE

The exact MLE can be formulated as a standard convex optimization problem,

$$\begin{aligned}&\text{minimize} \quad -Q = -\sum_{i=1}^{N}\sum_{j=0}^{s_i-1}\log(x_i-j),\\ &\text{subject to} \quad \mathbf{AX}=\mathbf{F}, \quad \text{and} \quad \mathbf{X}\succeq\mathbf{S},\end{aligned} \tag{117}$$

where $\mathbf{X}\succeq\mathbf{S}$ is a compact representation for $x_i \geq s_i, 1\leq i\leq N$.

This optimization problem can be solved by a variety of standard methods such as the Newton's method (Boyd and Vandenberghe, 2004, Chapter 10.2). Note that we can ignore the implicit inequality constraints, $\mathbf{X}\succeq\mathbf{S}$, if we start with feasible (i.e., satisfying both equality and inequality constraints) initial guess.

It turns out that the formulation in (117) will encounter numerical problems due to the inner summation in the objective function Q. Strictly speaking, we should use the integer programming algorithms because all variables are supposed to be integers. We formulate it approximately as a convex programming without worrying the integer constraints. Smoothing will bring in more numerical difficulties. Recall in estimating two-way associations we do not have this problem, because we have eliminated the summation in the objective function, by using the (integer) updating formula. In multi-way associations, we do not see any easy way to reformulate the objective function Q in a similar updating form.

To avoid the numerical problems, a simple solution is to assume "sample-with-replacement," under which the conditional likelihood becomes

$$P(\mathbf{S}|D_s;\mathbf{X},r) \propto \prod_{i=1}^{N}\left(\frac{x_i}{D}\right)^{s_i} \propto \prod_{i=1}^{N}x_i^{s_i}, \tag{118}$$

and the $\log$ likelihood would be

$$\log P(\mathbf{S}|D_s;\mathbf{X},r) \propto Q_r = \sum_{i=1}^{N}s_i\log x_i. \tag{119}$$

Our MLE problem can then be reformulated as

$$\begin{aligned}&\text{minimize} \quad -Q = -\sum_{i=1}^{N}s_i\log x_i,\\ &\text{subject to} \quad \mathbf{AX}=\mathbf{F}, \quad \text{and} \quad \mathbf{X}\succeq\mathbf{S},\end{aligned} \tag{120}$$

which is again a standard convex optimization problem. To simplify the notation, we neglect the subscript '"r" because throughout the rest of this section we will be working with the "sample-with-replacement" version of Q.

We can compute the gradient ($\bigtriangledown Q$) and Hessian ($\bigtriangledown^2 Q$). The gradient is a vector of the first derivatives of Q with respect to x_i, for $1 \leq i \leq N$,

$$\bigtriangledown Q = \left[\frac{\partial Q}{\partial x_i}, 1 \leq i \leq N\right] = \left[\frac{s_1}{x_1}, \frac{s_2}{x_2}, ..., \frac{s_N}{x_N}\right]^{\mathrm{T}}, \tag{121}$$

where the superscript T indicates "transpose"as we always work with column vectors.

The Hessian is a matrix whose $(i,j)^{th}$ entry is the partial derivative $\frac{\partial^2 Q}{\partial x_i x_j}$, i.e.,

$$\bigtriangledown^2 Q = -\begin{bmatrix} \frac{s_1}{x_1^2} & 0 & \cdots & 0 \\ 0 & \frac{s_2}{x_2^2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \frac{s_N}{x_N^2} \end{bmatrix} = -\text{diag}\left[\frac{s_1}{x_1^2}, \frac{s_2}{x_2^2}, ..., \frac{s_N}{x_N^2}\right]. \tag{122}$$

The Hessian has a very simple diagonal form, implying that the Newton's method will be the best algorithm for solving the optimization problem. We implement the equality constrained Newton's method with feasible start and backtracking line search (Boyd and Vandenberghe, 2004, Algorithm 10.1). A key step in this algorithm is to solve for the Newton's step, $\triangle \mathbf{X}_{nt}$:

$$\begin{bmatrix} -\bigtriangledown^2 Q & \mathbf{A}^{\mathrm{T}} \\ \mathbf{A} & 0 \end{bmatrix} \begin{bmatrix} \triangle \mathbf{X}_{nt} \\ \text{dummy} \end{bmatrix} = \begin{bmatrix} \bigtriangledown Q \\ 0 \end{bmatrix}. \tag{123}$$

Since the Hessian $\bigtriangledown^2 Q$ is a diagonal matrix, solving for the Newton's step in (123) can be speeded up substantially (e.g., using the block matrix inverse formula).

How do we obtain a feasible initial starting value, $\mathbf{X}_{ini}$? It is easy without the inequality constraints in (120). For example, it appears that a nice choice of the initial guess would be: $\hat{\mathbf{X}}_{ini} = \hat{\mathbf{X}}_{MF} - \mathbf{A}^{\mathrm{T}}(\mathbf{A}\mathbf{A}^{\mathrm{T}})^{-1}(-\mathbf{F} + \mathbf{A}\hat{\mathbf{X}}_{MF})$, which satisfies $\mathbf{A}\mathbf{X} = \mathbf{F}$ and minimizes the 2-norm $\|\hat{\mathbf{X}}_{MF} - \hat{\mathbf{X}}_{ini}\|_2^2$. Unfortunately, this choice of $\hat{\mathbf{X}}_{ini}$ does not satisfy the inequality constraint, $\mathbf{X} \succeq \mathbf{S}$, hence not particularly useful.

Our approach is to solve for the feasible initial guess from a simpler quadratic optimization problem,

$$\begin{aligned} &\text{minimize } \|\hat{\mathbf{X}}_{MF} - \mathbf{X}_{ini}\|_2^2 \\ &\text{subject to } \mathbf{A}\mathbf{X}_{ini} = \mathbf{F}, \text{ and } \mathbf{X}_{ini} \succeq \mathbf{S}. \end{aligned}$$

Alternatively, one can use the 1-norm, which leads to a standard linear programming problem. Both quadratic programming and linear programming can be easily solved using standard software (e.g., Matlab).

In fact, the whole convex optimization problem in (120) can be solved by numerical packages, although most of these packages may still require the user input of $\bigtriangledown Q$ and $\bigtriangledown^2 Q$.

We provide a sample implementation of the multi-way association estimator in Appendix B.

11.5 Variance Estimation

We will again apply the large sample theory to estimate the variance of the MLE, which will be a covariance matrix for multi-way associations. Recall that we have $N = 2^m$ variables and $m+1$ constraints. The effective number of variables would be $2^m - (m+1)$, which is also the dimension of the covariance matrix.

One approach will convert $\mathbf{X}$ into $\mathbf{Z}$, which is a vector of length $N-(m+1)$. For example (Boyd and Vandenberghe, 2004, Chapter 10.1.2), one can write $\mathbf{X} = \mathbf{B}\mathbf{Z} + \tilde{\mathbf{X}}$, where $\tilde{\mathbf{X}}$ is particular solution satisfying the margin constraints and could be treated as a constant; $\mathbf{B}$, a matrix of size $N \times (N-(m+1))$, is the null space of $\mathbf{A}$, i.e., $\mathbf{AB} = \mathbf{0}$. With this type of transformation, one can first estimate the variance in the $\mathbf{Z}$-space and then convert it back to the $\mathbf{X}$-space. Our approach, however, is simpler, by exploiting the specific structure of $\mathbf{A}$.

We seek a partition of $\mathbf{A} = [\mathbf{A_1}, \mathbf{A_2}]$, such that $\mathbf{A_2}$ is invertible. We may have to switch the columns of $\mathbf{A}$ in order to find an invertible $\mathbf{A_2}$. In our construction, the jth column of $\mathbf{A}_2$ is the column of $\mathbf{A}$ such that last entry of the jth row of $\mathbf{A}$ is 1. An example for $m = 3$ would be

$$\mathbf{A_1^{(3)}} = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \quad \mathbf{A_2^{(3)}} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}, \tag{124}$$

where $\mathbf{A_1}$ is the [1 2 3 5] columns of $\mathbf{A}$ and $\mathbf{A_2}$ is the [4 6 7 8] columns of $\mathbf{A}$. We can see that $\mathbf{A}_2$ constructed this way is always invertible because its determinant is always one.

Corresponding to the partition of $\mathbf{A}$, we partition $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2]^{\mathrm{T}}$. For example, when $m = 3$, $\mathbf{X}_1 = [x_1, x_2, x_3, x_5]^{\mathrm{T}}$, $\mathbf{X}_2 = [x_4, x_6, x_7, x_8]^{\mathrm{T}}$. We can then express $\mathbf{X}_2$ to be

$$\mathbf{X}_2 = \mathbf{A}_2^{-1}\left(\mathbf{F} - \mathbf{A}_1\mathbf{X_1}\right) = \mathbf{A}_2^{-1}\mathbf{F} - \mathbf{A}_2^{-1}\mathbf{A}_1\mathbf{X_1}. \tag{125}$$

The $\log$ likelihood function Q, which is separable, can then be expressed as

$$Q(\mathbf{X}) = Q_1(\mathbf{X}_1) + Q_2(\mathbf{X}_2). \tag{126}$$

By the matrix derivative chain rule, the Hessian of Q with respect to $\mathbf{X}_1$ would be

$$\bigtriangledown_1^2 Q = \bigtriangledown_1^2 Q_1 + \bigtriangledown_1^2 Q_2 = \bigtriangledown_1^2 Q_1 + \left(\mathbf{A}_2^{-1}\mathbf{A}_1\right)^{\mathrm{T}} \bigtriangledown_2^2 Q_2 \left(\mathbf{A}_2^{-1}\mathbf{A}_1\right), \tag{127}$$

where we use $\bigtriangledown_1^2$ and $\bigtriangledown_2^2$ to indicate the Hessians are with respect to $\mathbf{X}_1$ and $\mathbf{X}_2$, respectively.

Conditional on D_s, the Expected Fisher Information of $\mathbf{X}_1$ is

$$\begin{aligned} \mathrm{I}(\mathbf{X}_1) &= \mathrm{E}\left(-\bigtriangledown_1^2 Q\right) \\ &= -\mathrm{E}(\bigtriangledown_1^2 Q_1) + \left(\mathbf{A}_2^{-1}\mathbf{A}_1\right)^{\mathrm{T}} \mathrm{E}(-\bigtriangledown_2^2 Q_2) \left(\mathbf{A}_2^{-1}\mathbf{A}_1\right). \end{aligned} \tag{128}$$

Again, we approximate the expectations using the results from the margin-free case, i.e.,

$$\mathrm{E}(-\bigtriangledown_1^2 Q_1) = \mathrm{diag}\left[\mathrm{E}\left(\frac{s_i}{x_i^2}\right), x_i \in \mathbf{X}_1\right] \approx \frac{D_s}{D}\mathrm{diag}\left[\frac{1}{x_i}, x_i \in \mathbf{X}_1\right], \tag{129}$$

$$\mathrm{E}(-\bigtriangledown_2^2 Q_2) \approx \frac{D_s}{D}\mathrm{diag}\left[\frac{1}{x_i}, x_i \in \mathbf{X}_2\right]. \tag{130}$$

By the large sample theory, and also considering the finite population correction factor, we can approximate the (conditional) covariance matrix of $\mathbf{X}_1$ to be

$$\begin{aligned} \mathrm{Cov}(\mathbf{X}_1) &\approx \mathrm{I}(\mathbf{X}_1)^{-1}\frac{D - D_s}{D} \\ &\approx \left(\frac{D}{D_s} - 1\right)\left(\mathrm{diag}\left[\frac{1}{x_i}, x_i \in \mathbf{X}_1\right] + \left(\mathbf{A}_2^{-1}\mathbf{A}_1\right)^{\mathrm{T}} \mathrm{diag}\left[\frac{1}{x_i}, x_i \in \mathbf{X}_2\right]\left(\mathbf{A}_2^{-1}\mathbf{A}_1\right)\right)^{-1}. \end{aligned} \tag{131}$$

We can also use compute the Observed Fisher Information by not evaluating the expectations.

For a sanity check, we can verify that this approach recovers the same variance formula in the two-way association case. Recall that, when $m = 2$, we have

$$\bigtriangledown^2 Q = -\begin{bmatrix} \frac{s_1}{x_1^2} & 0 & 0 & 0 \\ 0 & \frac{s_2}{x_2^2} & 0 & 0 \\ 0 & 0 & \frac{s_3}{x_3^2} & 0 \\ 0 & 0 & 0 & \frac{s_4}{x_4^2} \end{bmatrix},\ \bigtriangledown_1^2 Q_1 = -\frac{s_1}{x_1^2},\ \bigtriangledown_2^2 Q_2 = -\begin{bmatrix} \frac{s_2}{x_2^2} & 0 & 0 \\ 0 & \frac{s_3}{x_3^2} & 0 \\ 0 & 0 & \frac{s_4}{x_4^2} \end{bmatrix} \tag{132}$$

$$\mathbf{A^{(2)}} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix},\ \mathbf{A_1^{(2)}} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix},\ \mathbf{A_2^{(2)}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \tag{133}$$

$$\begin{aligned} \left(\mathbf{A}_2^{-1}\mathbf{A}_1\right)^{\mathrm{T}} \bigtriangledown_2^2 Q_2 \mathbf{A}_2^{-1}\mathbf{A}_1 &= -\begin{bmatrix} 1 & 1 & -1 \end{bmatrix} \begin{bmatrix} \frac{s_2}{x_2^2} & 0 & 0 \\ 0 & \frac{s_3}{x_3^2} & 0 \\ 0 & 0 & \frac{s_4}{x_4^2} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ -1 \end{bmatrix} \\ &= -\frac{s_2}{x_2^2} - \frac{s_3}{x_3^2} - \frac{s_4}{x_4^2} \end{aligned} \tag{134}$$

Hence,

$$\begin{aligned} -\bigtriangledown_1^2 Q &= \frac{s_1}{x_1^2} + \frac{s_2}{x_2^2} + \frac{s_3}{x_3^2} + \frac{s_4}{x_4^2} \\ &= \frac{a_s}{a^2} + \frac{b_s}{(f_1 - a)^2} + \frac{c_s}{(f_2 - a)^2} + \frac{d_s}{(D - f_1 - f_2 + a)^2}, \end{aligned} \tag{135}$$

which leads to the same Fisher Information and variance for the two-way association case as we have already derived.

11.6 Unconditional Variance

Similar to two-way associations, the unconditional variance of the proposed MLE can be estimated by replacing $\frac{D}{D_s}$ in (131) with $\mathrm{E}\left(\frac{D}{D_s}\right)$, i.e.,

$$\mathrm{Cov}(\mathbf{X}_1)_{uc} \approx \left(\mathrm{E}\left(\frac{D}{D_s}\right) - 1\right)\left(\mathrm{diag}\left[\frac{1}{x_i}, x_i \in \mathbf{X}_1\right] + \left(\mathbf{A}_2^{-1}\mathbf{A}_1\right)^{\mathrm{T}} \mathrm{diag}\left[\frac{1}{x_i}, x_i \in \mathbf{X}_2\right]\left(\mathbf{A}_2^{-1}\mathbf{A}_1\right)\right)^{-1}. \tag{136}$$

Following the similar analysis as in two-way associations, we can get the approximate formulas

$$\mathrm{E}\left(\frac{D_s}{D}\right) \approx \min\left(\frac{k_1}{f_1}, \frac{k_2}{f_2}, ..., \frac{k_m}{f_m}\right), \tag{137}$$

$$\mathrm{E}\left(\frac{D}{D_s}\right) \approx \max\left(\frac{f_1}{k_1}, \frac{f_2}{k_2}, ..., \frac{f_m}{k_m}\right). \tag{138}$$

Again, the approximation (137) will over-estimate $\mathrm{E}\left(\frac{D_s}{D}\right)$ and (138) will under-estimate $\mathrm{E}\left(\frac{D}{D_s}\right)$ hence also under-estimates the unconditional variance.

Table 8
The same four words as in Table 8 are used for evaluating multi-way associations. There are in total four three-way combinations and one four-way combination.

	Case #	Words	Co-occurrences
Three-way	Case 3-1	THIS & HAVE & HELP	4940
	Case 3-2	THIS & HAVE & PROGRAM	2575
	Case 3-3	THIS & HELP & PROGRAM	1626
	Case 3-4	HAVE & HELP & PROGRAM	1460
Four-way	Case 4	THIS & HAVE & HELP & PROGRAM	1316

11.7 Evaluation

We use the same four words as in Table 5 to evaluate the multi-way association algorithm. Since our results are theoretical, this evaluation is merely a quick sanity check. For four words, there are four different combinations of three-way associations and one four-way associations, as numbered in Table 8.

We will only present the results for estimating x_1 (i.e., a in two-way associations) for all cases. The evaluations for four three-way cases are presented in Figures 39, 40, 41 and 42. From these figures, we can see that the proposed MLE is unbiased and has lower MSE than the margin-free baseline (MF). As in the two-way case, smoothing helps MLE but still hurts MF in most cases. Also, the experiments verify that our approximate variance formula are fairly accurate.

Figure 43 presents the evaluation results for the four-way association case, including MSE, smoothing, variance and bias. The results are similar to the three-way case.

We have used the empirical $\mathrm{E}\left(\frac{D}{D_s}\right)$ to compute the unconditional variance. Figure 44 plots $\max\left(\frac{f_1}{k_1}, \frac{f_2}{k_2}, ..., \frac{f_m}{k_m}\right) / \frac{D}{D_s}$ for all cases. The figure indicates that using $\max\left(\frac{f_1}{k_1}, \frac{f_2}{k_2}, ..., \frac{f_m}{k_m}\right)$ to estimate $\mathrm{E}\left(\frac{D}{D_s}\right)$ is still fairly accurate when the sample size is reasonable.

Combining the results of two-way associations for the same four words, we can study the trend how the proposed MLE improve the MF baseline. Figure 45(a) suggests that, compared with the MF baseline, the improvements of the proposed MLE decreases monotonically as the order of associations increases. This observation is not surprising, because the degree of the freedom is $2^m - (m+1)$, increasing exponentially as the the order m increases. In order words, the effect of the margin constraints decreases as m increases.

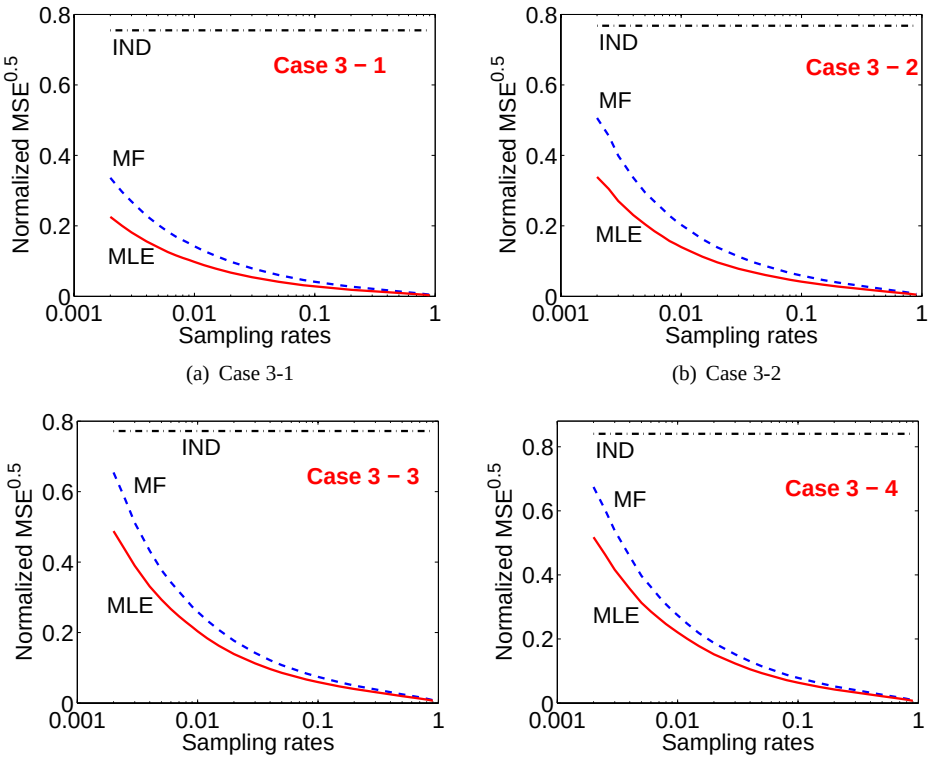
On the other hand, smoothing becomes more and more important as m increases, as shown in Figure 45(b), partly because of the data sparcity in high order associations.

12 Conclusion

We proposed a novel sketch-based procedure for constructing sample contingency tables. The method bridges two popular choices: (A) sampling over documents and (B) sampling over postings. Well-understood maximum likelihood estimation (MLE) techniques can be applied to sketches (or to traditional samples) to estimate word associations. We derived an exact cubic solution, $\hat{a}_{MLE}$, as well as a quadratic approximation, $\hat{a}_{MLE,a}$. The approximation is recommended because it is close to the exact solution, and easy to compute.

The proposed MLE methods were compared empirically and theoretically to a margin-free (MF) baseline, finding large improvements. When we know the margins, we ought to use them.

Not unsurprisingly, there is a trade-off between computational work (space and time) and statistical accuracy (variance or errors); reducing the sampling rate saves work, but costs accuracy. We derived formulas for variance, showing precisely how accuracy depends on sampling



(a) Case 3-1 (b) Case 3-2

(c) Case 3-3 (d) Case 3-4

Figure 39

In terms of $\frac{\text{MSE}(x_1)^{0.5}}{x_1}$, the proposed MLE is consistently better than the margin-free baseline (MF), which is better than the independence baseline (IND), for four three-way association cases.

rate. Sampling methods become more and more important with larger and larger collections. At Web scale, sampling rates as low as 10^{-4} may suffice for "ordinary" words.

Our sketch construction generalized Broder's sketch. Our method is more flexible in that we do not require fixed sample sizes. Using the same storage for the samples, our method can improve Broder's algorithm by roughly 50%. The improvement over random projections is even larger (e.g., $80\% - 90\%$).

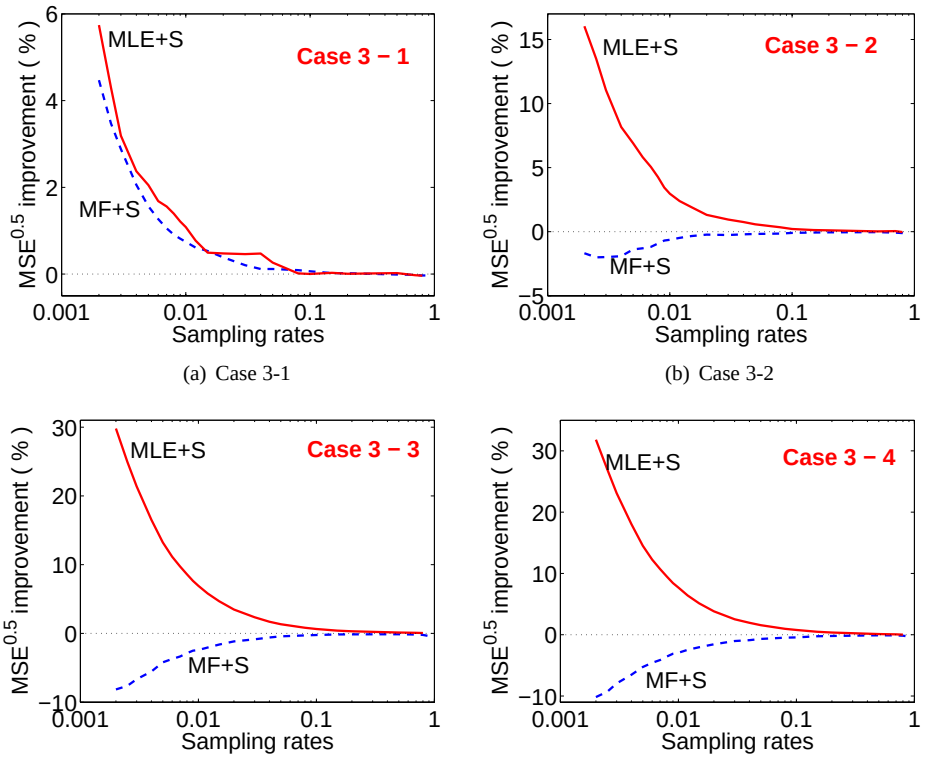
We have generalized the sampling algorithm and estimation method to multi-way associations, which is important for many applications such as estimating the number of page hits for a multi-word Web query.

Acknowledgments

The authors thank Chris Meek, David Heckerman, Mark Manasse, Robert Moore, Jonathan Goldstein, Trevor Hastie, David Siegmund, Art Owen, Robert Tibshirani, Bradley Efron and Andrew Ng.

A Sample C code for estimating two-way associations

```
#include <stdio.h>
#include <math.h>
#define MAX(x,y) ( (x) > (y) ? (x) : (y) )
#define MIN(x,y) ( (x) < (y) ? (x) : (y) )
```



(a) Case 3-1 (b) Case 3-2

(c) Case 3-3 (d) Case 3-4

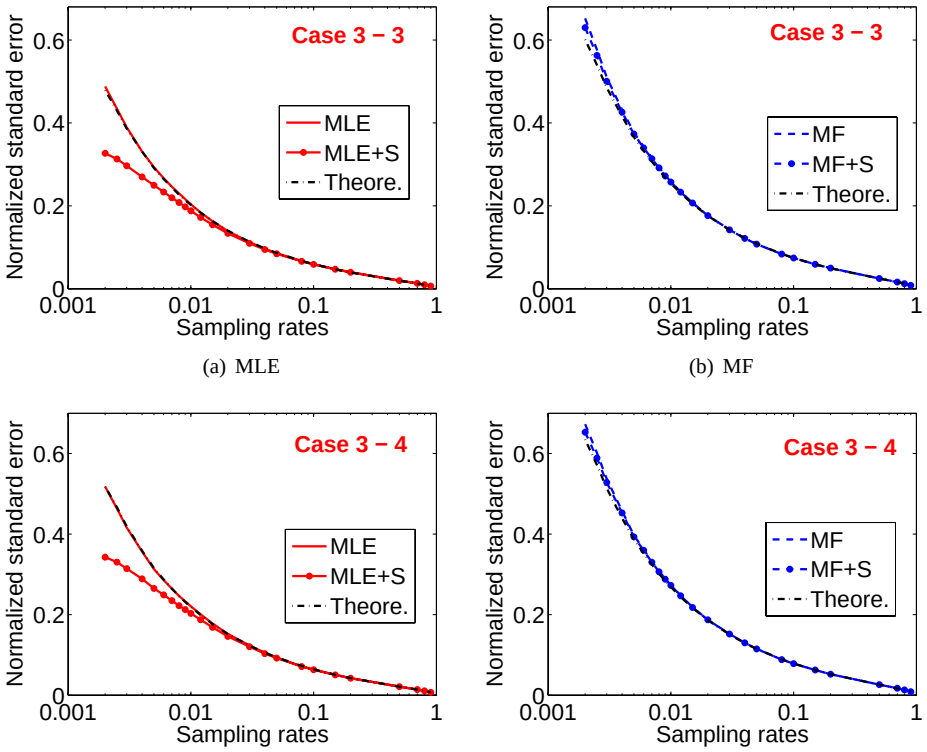
Figure 40
The simple "add-one" smoothing improves the estimation accuracies for the proposed MLE. Smoothing, however, in all cases except Case 3-1 hurts the margin-free estimator.

```
#define EPS 1e-10
#define MAX_ITER 50
int est_a_appr(int as,int bs,int cs, int f1, int f2);
int est_a_mle(int as,int bs, int cs, int ds, int f1, int f2,int D);

int main(void)
{
   int f1 = 10000, f2 = 5000, D = 65536;          // test data
   int as = 25, bs = 45, cs = 150, ds = 540;

   int a_appr = est_a_appr(as,bs,cs,f1,f2);
   int a_mle = est_a_mle(as,bs,cs,ds,f1,f2,D);
   printf("Estimate a_appr = %d\n",a_appr);       // output 1138
   printf("Estimate a_mle  = %d\n",a_mle);        // output 821
   return 0;
}

// The approximate MLE is the solution to a quadratic equation
int est_a_appr(int as,int bs,int cs, int f1, int f2)
{
   int sx = 2*as + bs,  sy = 2*as + cs, sz = 2*as+bs+cs;
   double  tmp = (double)f1*sy + (double)f2*sx;
   return  (int)((tmp-sqrt(tmp*tmp-8.0*f1*f2*as*sz))/sz/2.0);
}
```



(a) MLE (b) MF

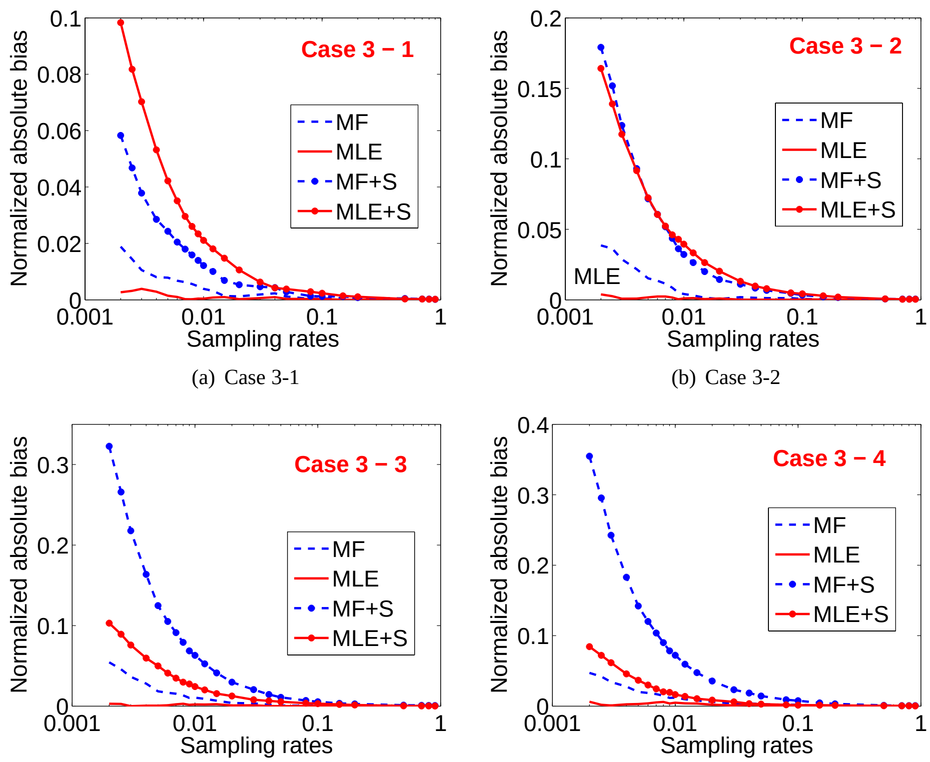
(c) MLE (d) MF

Figure 41

The proposed MLE is compared with the margin-free (MF) baseline in terms of $\frac{\mathrm{SE}(x_1)}{x_1}$, for Case 3-3 and Case 3-4 only. The proposed MLE has lower variance than MF. At low sampling rates, smoothing effectively reduces the variance for MLE but not for MF. The theoretical variance of MLE fits the empirical values very well. Note that we plug in the empirical E $\left(\frac{D}{D_s}\right)$ into (136) to estimate the unconditional variance. The errors due to this approximation are presented in Figure 44.

```
// Newton's method to solve for the exact MLE
int est_a_mle(int as,int bs, int cs, int ds, int f1, int f2,int D)
{
   int a_min = MAX(as,ds+f1+f2-D), a_max = MIN(f1-bs,f2-cs);
   int a1  = est_a_appr(as,bs,cs,f1,f2);  // A good start
   a1 = MAX( a_min, MIN(a1, a_max) );     // Sanity check

   int k = 0, a = a1;
   do {
      a = a1;
      double q = log(a+EPS) - log(a-as+EPS)
               +log(f1-a-bs+1+EPS) - log(f1-a+1+EPS)
               +log(f2-a-cs+1+EPS) - log(f2-a+1+EPS)
               +log(D-f1-f2+a+EPS) - log(D-f1-f2-ds+a+EPS);
      double dq = 1.0/(a+EPS)-1.0/(a-as+EPS)
                 -1.0/(f1-a-bs+1+EPS) + 1.0/(f1-a+1+EPS)
                 -1.0/(f2-a-cs+1+EPS) + 1.0/(f2-a+1+EPS)
                 -1.0/(D-f1-f2-ds+a+EPS) + 1.0/(D-f1-f2+a+EPS);
      a1 = (int)(a - q/dq);   a1 = MAX(a_min, MIN(a1,a_max));
      if( ++k > MAX_ITER )  break;
   }while( a1 != a );
```



(a) Case 3-1 (b) Case 3-2

(c) Case 3-3 (d) Case 3-4

Figure 42
The estimation biases, in terms of $\frac{|\text{bias}(x_1)|}{x_1}$, verify that our proposed ME is unbiased, unlike the margin-free baseline.

```
   return a;
}
```

B Sample Matlab Code for Estimating Multi-way Associations

```
function test_program
% %%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Authors: Ping Li  and  Kenneth Church     %%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% A short program for testing the multi-way association algorithm.
% First generate a random gold standard dataset. Then construct
% sketches by sampling a certain portion of the postings. Associations
% are estimated by the exact MLE as well as the margin-free (MF) method.
%
clear all;
m = max(2,ceil(rand*6));     % Number of words (random)
D = 1000*m;                  % Total number of documents
f = ceil(rand(m,1)*D/2);     % document frequencies (random)

P{1} = sort(randsample(D,f(1)));  % Posting of the first word (random)
Pc = setdiff(1:D, P{1})';         % Compliment of the posting

% The postings of words 2 to m are randomly generated. 30% are
% sampled from the postings of word 1.
```

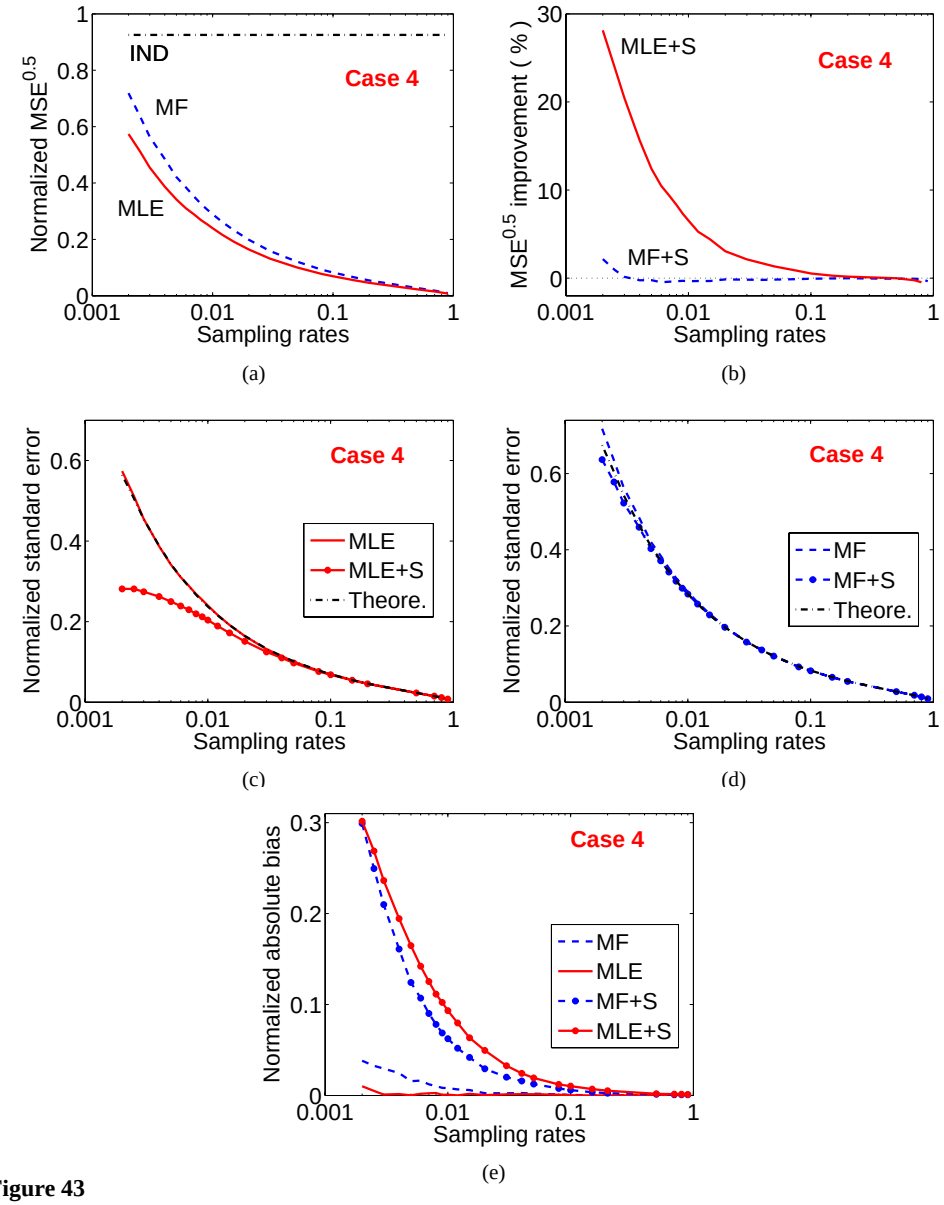


Figure 43
The evaluation results for the four-way case are presented in five sub-figures. (a): The proposed MLE has smaller MSE than the margin-free (MF) baseline, which has smaller MSE than the independence baseline. (b): Smoothing considerably improve the accuracy for MLE and also slightly improves MF. (c): For the proposed MLE, the theoretical prediction fits the empirical variance very well. Smoothing considerably reduces variance. (d): For the MF baseline, smoothing slightly reduces variance. (e): The MLE is unbiased while the MF baseline has slightly higher bias. Smoothing increases bias.

```
for i = 2:m
    k = ceil(0.3*min(f(i),f(1)));
    P{i} = sort([randsample(P{1},k);randsample(Pc,f(i)-k)]);  % Postings
end
X = compute_intersection(P,D);     % Gold standard associations
```

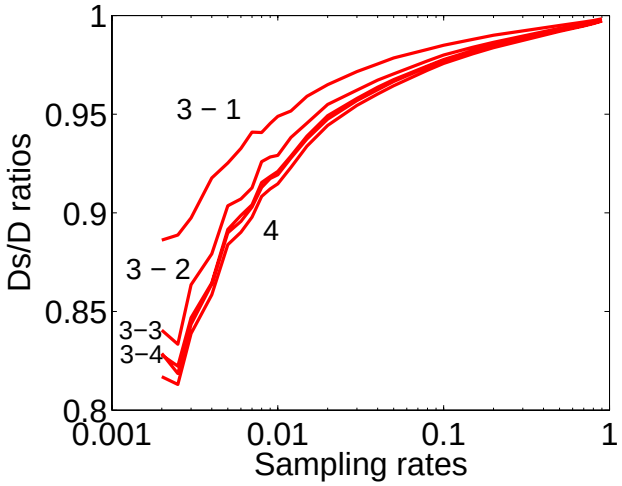


Figure 44

The ratios $\max\left(\frac{f_1}{k_1}, \frac{f_2}{k_2}, ..., \frac{f_m}{k_m}\right) / \frac{D}{D_s}$ are plotted for all cases. At sampling rates > 0.01, the ratios are $> 0.9 - 0.95$, indicating good accuracy.

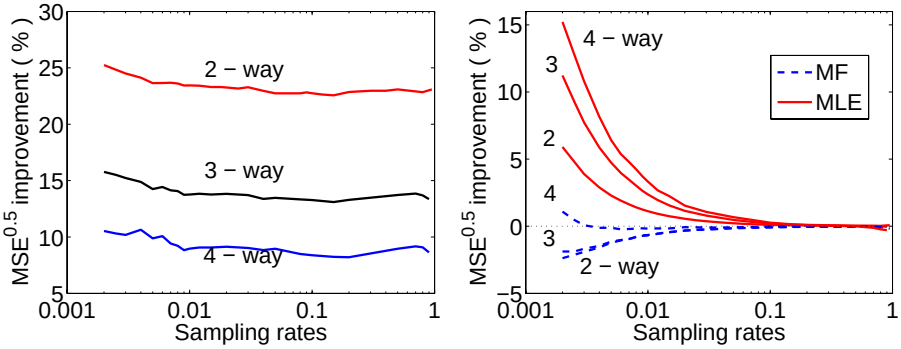


Figure 45

(a): Combining the three-way, four-way, and the two-way association results for the four words in the evaluations, the average relative improvements of $MSE^{0.5}$ suggests that the proposed MLE is consistently better the MF baseline but the improvement decreases monotonically as the order of associations increases. (b): Average $MSE^{0.5}$ improvements due to smoothing imply that smoothing becomes more and more important as the order of association increases.

```
pc = 1;  % Pseudo-count(pc), pc=0 for no smoothing, pc=1 for "add-one".
sampling_rate = 0.1;
for i = 1:m
    k = ceil(sampling_rate*f(i));
    K{i} = P{i}(1:k);              % Sketches
end
% Estimate the associations and covariance matrices
[X_MLE, X_MF, Var_c, Var_o] = multi_way_est(K,f,D,pc);
% Display the estimations of associations
[X X_MLE X_MF]             %  [Gold standard, MLE, MF]
```

```
function [X_MLE, X_MF, Var_c, Var_o] = multi_way_est(K,f,D,pc);
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

```
% Authors:  Ping Li  and Kenneth Church         %%%%%%%%%%%%%%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
% Matlab code for estimating multi-way associations
% K:      Sketches   (Cell array data type)
% f:      Document frequencies, a column vector
% D:      Total number of documents
% pc:     Pseudo-count for smoothing.
% X_MLE:  Maximum likelihood estimator (MLE),  a column vector
% X_MF :  Margin-free (MF) estimator, a column vector
% Var_c:  Conditional (on Ds) covariance matrix, using the estimated X,
% Var_o:  Covariance computed using the observed Fisher information
%
pc = max(pc,1e-4); % Always use a small pc for numerical stability.
m = length(K);     % The order of associations, i.e., number of words.
[A,A1,A2,A3,ind1,ind2] = gen_A(m);   % Margin constraint matrix

for i = 1:m;
    last_elem(i) = K{i}(end);
end
Ds = min(last_elem);
for i = 1:m
    K{i} = K{i}(find(K{i}<=Ds));  % Trim sketches according to D_s
end

S = compute_intersection(K,Ds);   % Intersect the sketches to get samples
[X_MLE, X_MF] = newton_est(pc,S,Ds,D,A,f); % Estimate X

% Conditional variance
Z_c = 1./(X_MLE+eps); Z1_c = diag(Z_c(ind1)); Z2_c = diag(Z_c(ind2));
Var_c = inv(Z1_c + A3'*Z2_c*A3)*(D/Ds-1);

% Observed variance
Z_o = S./(X_MLE+eps).^2; Z1_o = diag(Z_o(ind1)); Z2_o = diag(Z_o(ind2));
Var_o = inv(Z1_o + A3'*Z2_o*A3)*(D-Ds)/D;
```

```
function [X_MLE,X_MF] = newton_est(pc,S,Ds,D,A,f)
% Estimate multi-way associations by solving a convex
% optimization problem using the Newton's method.

NEWTON_ERR = 0.001;        % Threshold for termination.
MAX_ITER = 50;             % Maximum allowed iteration.
N = length(S); m = length(f);  F = [f;D];
pc = min(pc,(D-Ds)/N);  % Adjust pc, if Ds is close to D.

% Solve a quadratic programming problem to find an initial
% guess of the MLE that minimizes the 2-norm with respect to
% the MF estimation and satisfies the constraints.
while(1)
    X_MF = (S+pc)./(Ds+N*pc)*D;      % Margin-free estimations.
    [X0,dummy,flag] = quadprog(2*eye(2^m),-2*X_MF,[],[],A,F,S+pc);
    if(flag == 1)    break;    end
    pc = pc/2;  % Occasionally need reduce pc for a feasible solution.
end

S = S + pc;  X_MLE = X0;  iter = 0;
while(1);
    D1 = -S./(X_MLE+eps);         % Gradient  (first derivatives)
    D2 = diag(S./(X_MLE.^2+eps)); % Hessian   (second derivatives)

    % Solve a linear system of equations for the Newton's step.
```

```
    M = [D2 A'; A zeros(size(A,1),size(A,1))];
    dx = M\[-D1; zeros(size(A,1),1)]; dx = dx(1:size(D2,1));

    lambda = (dx'*D2*dx)^0.5;       % Measure of errors
    iter = iter + 1;
    if(iter>MAX_ITER | lambda^2/2<NEWTON_ERR)     break;   end

    % Backtracking line search for a good Newton step size.
    z = 1; Alpha = 0.1; Beta = 0.5;     iter2 = 0;
    while(min(X_MLE+z*dx-S)<0|S'*log(X_MLE./(X_MLE+z*dx))>=Alpha*z*D1'*dx);
        if(iter2 >= MAX_ITER)    break;   end
        z = Beta*z;  iter2 = iter2 + 1;
    end
    X_MLE = X_MLE + z*dx;
end
```

```
function  S = compute_intersection(K,Ds);
% Compute the intersections to generate a table with N = 2^m
% cells. The cells are ordered in terms of the binary representation
% of integers from 0 to 2^m-1, where m is the number of words.

m = length(K);  bin_rep = char(dec2bin(0:2^m-1)); S = zeros(2^m,1);
for i = 0:2^m-1;
    if(bin_rep(i+1,1) == '0')
        c{i+1} = K{1};
    else
        c{i+1} = setdiff([1:Ds]',K{1});
    end
    for j = 2:m
        if(bin_rep(i+1,j) == '0')
            c{i+1} = intersect(c{i+1},K{j});
        else
            c{i+1} = setdiff(c{i+1},K{j});
        end
    end
    S(i+1) = length(c{i+1});
end
```

```
function [A,A1,A2,A3,ind1,ind2] = gen_A(m)
% Generate the margin constraint matrix and compute its decompositions
% for analyzing the covariance matrix

t1 = num2str(dec2bin(0:2^m-1));  t2  = zeros(2^m,m*2-1);
t2(:,1:2:end) = t1;  t2(:,2:2:end) = ',';
A = xor(str2num(char(t2))',1); A = [A;ones(1,2^m)];

for i = 1:size(A,1);
    [last_one(i)] = max(find(A(i,:)==1));
end
ind1 = setdiff((1:size(A,2)),last_one); ind2 = last_one;
A1 = A(:,ind1); A2 = A(:,ind2);  A3 = inv(A2)*A1;
```

References

Alan Agresti. 2002. *Categorical Data Analysis*. John Wiley & Sons, Inc., Hoboken, NJ, second edition.

Ricardo Baeza-Yates and Berthier Ribeiro-Neto. 1999. *Modern Information Retrieval*. ACM Press, New York, NY.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57(1):289–300.

Stephen Boyd and Lieven Vandenberghe. 2004. *Convex Optimization*. Cambridge University Press, Cambridge, UK. Also online http://www.stanford.edu/ boyd/bv_cvxbook.pdf.

Sergey Brin and Larry Page. 1998. The anatomy of a large-scale hypertextual web search engine. In *Proceedings of the Seventh International World Wide Web Conference*, pages 107–117, Brisbane, Australia, April.

Andrei Z. Broder, Steven C. Glassman, Mark S. Manasse, and Geoffrey Zweig. 1997. Syntactic clustering of the web. In *Proceedings of the 6th International Conference on World Wide Web*, pages 1157 – 1166, Santa Clara, California, April.

Andrei Z. Broder, Moses Charikar, Alan M. Frieze, and Michael Mitzenmacher. 1998. Min-wise independent permutations (extended abstract). In *Proceedings of 30th STOC*, pages 327–336, Dallas, Texas, May.

Andrei Z. Broder, Moses Charikar, Alan M. Frieze, and Michael Mitzenmacher. 2000. Min-wise independent permutations. *Journal of Computer Systems and Sciences*, 60(3):630–659.

Andrei Z. Broder. 1997. On the resemblance and containment of documents. In *Proceedings of the Compression and Complexity of Sequences*, pages 21–29, Positano, Italy, June.

Moses S. Charikar. 2002. Similarity estimation techniques from rounding algorithms. In *Proceedings of the thirty-fourth annual ACM symposium on Theory of computing*, pages 380–388, Montreal, Quebec, Canada, May.

Stanley F. Chen and Joshua Goodman. 1996. An empirical study of smoothing techniques for language modeling. In *Proceedings of ACL*, pages 310–318, Santa Cruz, CA.

Stanley F. Chen and Joshua Goodman. 1998. An empirical study of smoothing techniques for language modeling. Technical Report TR-10-98, Center for Research in Computing, Harvard University, MA.

Yuan S. Chow and Herbert Robbins. 1965. On the asymptotic theory of fixed-width sequential confidence intervals for the mean. *The Annals of Mathematical Statistics*, 36(2):457–462.

Kenneth W. Church and William A. Gale. 1991. A comparison of the enhanced good-turing and deleted estimation methods for estimating probabilities of english bigrams. *Computer, Speech and Language*, 5(1):19–54.

Kenneth Church and Patrick Hanks. 1991. Word association norms, mutual information and lexicography. *Computational Linguistics*, 16(1):22–29.

Thomas M. Cover and Joy A. Thomas. 1991. *Elements of Information Theory*. John Wiley & Sons, Inc., New York, NY.

Scott Deerwester, Susan T. Dumais, George W. Furnas, and Thomas K. Landauer. 1999. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1):61–74.

Richard Durrett. 1995. *Probability: Theory and Examples*. Duxbury Press, Belmont, CA, second edition.

Efron and D.V. Hinkley. 1978. Assessing the accuracy of the maximum likelihood estimator: Observed versus expected fisher information. *Biometrika*, 65:457–487.

Bradley Efron and Robert Tibshirani. 2002. Empirical bayes methods and false discovery rates for microarrays. *Genetic Epidemiology*, 23(1):70–86.

Hector Garcia-Molina, Jeffrey D. Ullman, and Jennifer Widom. 2002. *Database Systems: the Complete Book*. Prentice Hall, New York, NY.

Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. 2004. *Bayesian Data Analysis*. Chapman & Hall/CRC, New York, NY, second edition.

Michel X. Goemans and David P. Williamson. 1995. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Journal of the Association for Computing Machinery*, 42(6):1115–1145.

Irving John Good. 1953. The population frequencies of species and the estimation of population parameters. *Biometrika*, 40(16):237–264.

Albert Sydney Hornby, editor. 1989. *Oxford Advanced Learner's Dictionary of Current English*. Oxford University Press, Oxford, UK, fourth edition.

Piotr Indyk and Rajeev Motwani. 1998. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *Proceedings of 30th STOC*, pages 604–613, Dallas, Texas, May.

Slava M. Katz. 1987. Estimation of probabilities from sparse data for the language model component of a speech recognizer. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 35(3):400–401.

Erich L. Lehmann and George Casella. 1998. *Theory of Point Estimation*. Springer, New York, NY, second edition.

Ping Li, Debashis Paul, Ravi Narasimhan, and John Cioffi. 2005. On the asymptotic and approximate distribution of SINR for the linear MMSE MIMO receiver. Technical Report 2005-15, Department of Statistics, Stanford University, CA, June.

Chris D. Manning and Hinrich Schutze. 1999. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, MA.

Robert C. Moore. 2004. On log-likelihood-ratios and the significance of rare events. In Dekang Lin and Dekai Wu, editors, *Proceedings of EMNLP 2004*, pages 333–340, Barcelona, Spain, July. Association for Computational Linguistics.

Arthur Nadas. 1969. An extension of a theorem of Chow and Robbins on sequential confidence intervals. *The Annals of Mathematical Statistics*, 40(2):667–671.

Judy Pearsall, editor. 1998. *The New Oxford Dictionary of English*. Oxford University Press, Oxford, UK.

Deepak Ravichandran, Patrick Pantel, and Eduard Hovy. 2005. Randomized algorithms and nlp: Using locality sensitive hash function for high speed noun clustering. In *Proceedings of ACL*, pages 622–629, Ann Arbor, June.

Sheldon M. Ross. 1996. *Stochastic Processes*. John Wiley & Sons Inc, New York, NY, second edition.

Sheldon Ross. 2002. *A First Course in Probability*. Prentice-Hall, Upper Saddle River, NJ, sixth edition.

Gerard Salton. 1989. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley, New York, NY.

Jun Shao. 1999. *Mathematical Statistics*. Springer, New York, NY, first edition.

David Siegmund. 1985. *Sequential Analysis, Tests and Confidence Intervals*. Springer-Verlag, New York, NY.

Kyle Siegrist. 1997. *Finite Sampling Models, http://www.ds.unifi.it/VL/VL_EN/urn/index.html*. Virtual Laboratories in Probability and Statistics, Huntsville, AL.

John D. Storey. 2003. The positive false discovery rate: a bayesian interpretation and the q-value. *Annals of Statistics*, 31(6):2013–2035.

Yee W. Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. 2004. Hierarchical dirichlet processes. Technical Report 653, Department of Statistics, University of California, Berkeley, CA.

William N. Venables and Brian D. Ripley. 2002. *Modern Applied Statistics with S*. Springer-Verlag, New York, NY, fourth edition.

Eric W. Weisstein. 2005a. *Binet's Log Gamma Formulas, http://mathworld.wolfram.com/BinetsLogGammaFormulas.html*. MathWorld–A Wolfram Web Resource.

Eric W. Weisstein. 2005b. *Cubic Formula, http://mathworld.wolfram.com/CubicFormula.html*. MathWorld–A Wolfram Web Resource.

Ian H. Witten, Alstair Moffat, and Timothy C. Bell. 1999. *Managing Gigabytes: Compressing and Indexing Documents and Images*. Morgan Kaufmann Publishing, San Francisco, CA, second edition.