The Computer Journal Advance Access published May 14, 2015

doi:10.1093/comjnl/bxv031

Harmony Assumptions in Information Retrieval and Social Networks

Thomas Roelleke[1,*], Andreas Kaltenbrunner[2] and Ricardo Baeza-Yates[3]

[1] *Queen Mary University of London, London, UK*
[2] *Barcelona Media, Barcelona, Spain*
[3] *Yahoo Labs Barcelona, Barcelona, Spain*
**Corresponding author: thor@eecs.qmul.ac.uk*

In many applications, *independence* of event occurrences is assumed, even if there is evidence for dependence. Capturing dependence leads to complex models, and even if the complex models were superior, they fail to beat the simplicity and scalability of the independence assumption. Therefore, many models assume independence and apply heuristics to improve results. Theoretical explanations of the heuristics are seldom given or generalizable. This paper reports that some of these heuristics can be explained as encoding dependence in an exponent based on the *generalized harmonic sum*. Unlike independence, where the probability of subsequent occurrences of an event is the product of the single event probability, *harmony* is based on a product with decaying exponent. For independence, the sequence probability is $p^{1+1+\cdots+1} = p^n$, whereas for harmony, it is $p^{1+1/2+\cdots+1/n}$. The generalized harmonic sum leads to a spectrum of *harmony assumptions*. This paper shows that harmony assumptions naturally extend probability theory. An experimental evaluation for information retrieval (IR; term occurrences) and social networks (SN's; user interactions) shows that assuming harmony is more suitable than assuming independence. The potential impact of harmony assumptions lies beyond IR and SN's, since many applications rely on probability theory and apply heuristics to compensate the independence assumption. Given the concept of harmony assumptions, the dependence between multiple occurrences of an event can be reflected in an intuitive and effective way.

Keywords: probability theory; modelling of dependence assumptions; harmonic sum; generalized binomial probability; information retrieval; social networks

Received 31 March 2014; revised 23 January 2015
Handling editor: Joemon Jose

1. INTRODUCTION

In a variety of applications, the independence assumption is employed for keeping probabilistic reasoning simple and scalable. Even though event occurrences may be known to be dependent, it often appears to be non-realistic to capture explicitly the dependence. Instead, parameters are added to fix the error made by assuming independence.

Parameter tuning in information retrieval (IR) models has led to 'heuristics' that are known to be conducive to good retrieval quality. Most important for retrieval quality is the term frequency (TF) quantification of the BM25 ranking formula [1, 2]. To illustrate what the BM25-TF quantification is about, consider an event (term) t, and let p_t be the single event probability. Then, in basic probability theory, for a sequence with n_t occurrences of event t, the sequence probability is $p_t^{n_t}$. The exponent being the total count means to assume *independence*. The BM25-TF can be interpreted as estimating the sequence probability differently, namely as $p_t^{2 \cdot n_t/(n_t+1)}$. What type of dependence does this sequence probability reflect?

The work reported in this paper is the result of a systematic process to explain the type of dependence assumption that is inherently modelled by TF quantifications applied in IR. The motivation of theoretical research is that if we were able to analytically describe the dependence, then we could arrive at a probability theory that helps to replace the 'heuristics'. At the same time, the achievements regarding IR ranking quality potentially can be generalized and transferred to other application domains that rely on probability theory.

The wider picture of this research can be related to [3]. The preface describes a scene where dad and daughter enter a cave.

- 'Dad, that boulder at the entrance, if it comes down, we are locked in'.
- 'Well, it stood there the last 10 000 years, so it won't fall down just now'.
- 'Dad, will it fall down one day?'
- 'Yes'.
- 'So it is more likely to fall down with every day it did not fall down?'

This scenario elicits the conflict arising when considering evidence, let it be for the occurrence or non-occurrence of an event. How is the probability of subsequent occurrences of an event affected by observed occurrences? This paper proposes and investigates 'harmony assumptions' for modelling the probabilities of subsequent event occurrences. For harmony, the probability of subsequent event occurrences is greater than for the early occurrences.

The notion of harmony encompasses and enriches the traditional way assumptions are modelled and referred to. It leads to a versatile probability theory with new terminology that allows for the intuitive formulation of scalable assumptions in a continuous spectrum from disjointness over independence to subsumption.

1.1. Structure of this paper

The first part discusses the *Background* (Section 2). We revisit independence vs dependence in IR and Social Networks (SN's), before engaging into a technical introduction of the main concepts that underpin this paper: *Probability Theory* (Independence Assumption, Section 2.4) and *IR Models* (TF-IDF and BM25, Section 2.5).

The second part introduces the *Harmony Assumptions* (Section 4). This is followed by *Sequence Probabilities* (Section 5) and *Frequency Probabilities* (Section 6). For frequency probabilities, we define the *harmonic binomial probability* (Section 6.3), a variant of the generalized binomial probability (Section 6.2) that is based on the harmonic sum.

The third part reports the *Analytical Verification* (Section 7) and the *Experimental Evaluation* (Section 8). We measure the dependencies among word occurrences in documents (IR scenario) and user interactions with recipients (SN scenario). It came as a surprise to find a similar type of harmony in both scenarios.

The appendix contains technical details.

2. BACKGROUND

There is a wide range of research aiming at solutions to the problem of 'modelling of event dependencies'. We review first the work that relates probability theory, randomness, IR and SN. Then, we provide the background of probability theory and IR models required to position the role of harmony assumptions.

2.1. Independence vs dependence in IR

The issue that the occurrences of a term are dependent is reflected by Zipf's law. How is the dependence related to Poisson? Early work to capture the dependence includes 'N-Poisson document modelling' [4], 'Poisson mixtures' [5] and 'IDF: Deviation from Poisson' [6]. The overall conclusion is that the independence assumption underlying the Poisson distribution does not reflect the actual distribution of terms (term frequencies), and that in the most widely known IR model, namely TF-IDF, there is somehow an inherent way to capture the dependence.

The two seminal papers 'Some Simple Approximations of the 2-Poisson Model' [7] and 'Okapi at TREC' [1, 2] brought the most effective TF quantification, the BM25-TF '$\mathrm{tf}_d/(\mathrm{tf}_d + K_d)$', where tf_d is the frequency of term t in document d, and K_d is a normalization parameter for a document length pivotization [8].

Some basic transformations (see Section 2.5.5) show that the BM25-TF can be related to an expression of the form $2 \cdot n/(n+1)$. This is the convergence value of the harmonic sum of Gaussian sums (see Section 4), and this started the research on harmony assumptions.

Another important TF quantification is the logarithmic TF, $\log(\mathrm{tf}_d + 1)$, common in text classification [9–12]. The logarithmic TF count ('ltc') is crucial for achieving good quality.

Bending the total count, like the BM25-TF and logarithmic TF do, coincides with the overall evidence that a Poisson-based model assigns too little of the probability mass to possible worlds with several occurrences of the same event. [13] 'TF normalization via Pareto distributions' and [14] 'Divergence From Randomness (DFR)' point at ways to measure the divergence between randomness and observed probabilities, and exploit these for ranking. However, there is no explicit coverage of the type of dependence.

Tuning the TF quantification can be considered as a research direction in IR. The work circles around Dirichlet priors, DFR, multinomial distributions and dependencies. For example: 'Dirichlet priors for TF normalizations' [15], 'TF Normalization Tuning for BM25 and DFR' [16], 'Retrieval based on Dirichlet Compound Multinomial Distribution' [17] and 'High-Order Word Dependence Mining' [18]. This work shows that the TF quantification and the coverage of dependencies are important and conducive to retrieval quality. Many of the models are complex, and although they capture the dependence (e.g. conditional probabilities), there is no analytical model. It is difficult to generalize the results to other application domains.

The late 90s saw language modelling (LM) [19] becoming popular in IR. This revived research on explaining roots and finding semantics for TF-IDF [20]. 'Approaches for using probabilistic dependencies' [21] provides an account of the techniques applied in IR (language modelling).

The research regarding ‘event spaces’ [22] and dualities between IR models [23, 24] produced a range of insights. This pathway led to the notion ‘semi-subsumption’ [25], an assumption lying between independence and subsumption.

The modelling of event dependencies is also a research topic in probabilistic DB’s [26, 27]. The area of probabilistic databases sparkled a wide range of issues regarding the coverage of dependencies, and the safe and efficient processing of queries: [28–34]. The research in probabilistic DB’s concerns the modelling of event dependencies to obtain ‘correct’ probabilities generated by ‘safe’ relational algebra expressions. The efficient processing is a major challenge when following the way traditional probability theory captures dependencies.

2.2. Independence vs dependence in SN’s

Modelling of human interaction in SN’s traditionally does not assume independence. One of the most applied models for human interaction is the preferential attachment model [35], which is based on the assumption that the likelihood to interact with a person is a function of the number of times one has previously interacted with this person. Preferential attachment leads to power-law (PL) distributions [36].

Nevertheless, in the IR-related task of link prediction, (i.e. estimate the likelihood of existence of a link between two nodes, based on observed links and the attributes of nodes) [37], most studies only focus on undirected and unweighted networks [38]. Connection weights (e.g. the number of interactions between friends) have been taken into account in [39] as linear sums. This idea has been extended in [40] using exponents in the weights. Interestingly, the best results are obtained in some cases for negative exponents, indicating a close relationship with the harmony assumption we propose in this paper.

2.3. Independence vs dependence before 1990

Research on uncertain reasoning and fuzzy theory suggest for long that the independence assumption needs alternatives. Fuzzy theory [41, 42] can be viewed as assuming events to be subsumed if the T-norm is based on the minimum and maximum of probabilities.

Early work addressing the independence and dependence of events include [43] ‘On the independence assumption underlying Bayesian updating’, [44] ‘A generalized term dependence model in IR’, [45] ‘An evaluation of term dependence models in IR’, [46, 47] ‘Independence assumptions and Bayesian updating’ and [48] ‘Boolean queries and term dependencies in probabilistic retrieval models’.

This surface overview pointing at the wide range of research underlines that the intuitive and analytical formulation of dependence is a long-standing problem. The fact that the TF quantification in IR reflects a dependence, combined with the fact that the TF quantification has a strong impact on retrieval quality, gives the ground for a general model of dependence.

2.4. The independence assumption

The independence assumption is simple, scales well and is applied in many scenarios. Let t_i denote an event, and $\Omega = \{t_1, \ldots, t_n\}$ the set of events. Let $d = (t_1, t_2, t_1, \ldots)$ be a sequence of event occurrences. Then, the independence assumption is

$$P(d) = \prod_{t \text{ IN } d} P(t) = \prod_{t \in d} P(t)^{n(t,d)} \qquad (1)$$

The notation ‘t IN d’ views d as a sequence of event occurrences, whereas ‘$t \in d$’ views d as a distinct set of events, and $n(t, d)$ is the number of occurrences of event t in sequence d. For example, for a document d with three words, the document probability is

$$P(d) = P(\text{sailing, boat, sailing}) = P(\text{sailing})^2 \cdot P(\text{boat})^1$$

There are two independence assumptions here. First, the different events, $t \in d$, are assumed to be independent. Secondly, the multiple occurrences of each event t are assumed to be independent. This paper addresses the second independence assumption, and proposes to make a harmony assumption.

2.5. IR models: TF-IDF and BM25

TF-IDF and BM25 are two of the main IR models [49, 50]. IR models have probabilistic roots, but contain heuristic components such as TF quantifications. TF-IDF and BM25 are the document-likelihood models, whereas LM is the query-likelihood model. For this paper, we focus on the document-likelihood side, since for the probability $P(d)$, the dependence matters. We briefly discuss the issues that arise when relating TF-IDF to probability theory, and where BM25 steps in. The aim is to highlight the relationship between TF quantifications and the independence assumption.

2.5.1. *TF: within-document term frequency*

Let t be a term and d a document. Let tf_d be the *term frequency* (total count) of term t in document d. The TF quantifications correspond to independence and dependence assumptions.

$$\begin{array}{ll} \text{TF}_{\text{total}}(t,d) := \text{tf}_d & \text{independence} \\ \text{TF}_{\log}(t,d) := \log(1+\text{tf}_d) & \text{what type of dependence?} \\ \text{TF}_{\text{BM25}}(t,d) := \dfrac{\text{tf}_d}{\text{tf}_d + K_d} & \text{what type of dependence?} \end{array}$$

$\text{TF}_{\log}$ is a quantification where the impact of high frequencies is less than for independence. TF_{BM25} assigns even less impact to high frequencies (K_d is discussed in Section 2.5.4, BM25). What type of dependence do $\text{TF}_{\log}$ and TF_{BM25} reflect? This paper shows analytically that both $\text{TF}_{\log}$ and TF_{BM25}, correspond to making a harmony assumption regarding the subsequent occurrences of an event.

2.5.2. *IDF: inverse document frequency*

Though this paper mainly addresses the dependence modelled by the TF, it is important to capture some of the IDF issues. Regarding the definition of IDF, let $\mathrm{df}(t, c)$ be the *document frequency* of term t in collection c; let $N_D(c)$ be the total number of *Documents* in collection c. Let $P_D(t \mid c) := \mathrm{df}(t, c)/N_D(c)$ denote the *Document-based* probability of term t in collection c. The common definition is

$$\mathrm{IDF}(t, c) := \log(1/P_D(t \mid c))$$

Without exploring details about theories on IDF, we briefly point out that there are three event spaces regarding the representation of the document event:

(1) binary vector: $\vec{d} = (1, 0, 1, 0, \ldots)$: $x_i \in \{0, 1\}$;
(2) frequency vector: $\vec{d} = (1, 0, 2, 0, \ldots)$: $f_i \in \{0, 1, 2, \ldots\}$;
(3) term sequence: $d = (t_3, t_1, t_3, \ldots)$: $t_i \in$ set of terms.

TF-IDF can be derived from any of the three spaces. (1) IDF is related to the *binary* independence retrieval model [51]. (2) IDF is related to the Poisson probability of the term *frequency* [4, 6, 52]. (3) TF-IDF is dual to LM, i.e. it can be derived based on a *term sequence* [20, 24]. The parallel derivation of models [23] distills the event spaces [22, 53]. Overall, we recall that IDF is based on the probability of a document. Therefore, the TF quantification models the dependencies between *document occurrences*.

2.5.3. *Probabilistic Root of TF-IDF*

Let d be a document, q a query, c a collection and t a term. The TF-IDF-based retrieval status value (RSV) is defined as follows:

$$\mathrm{RSV}_{\text{TF-IDF}}(d, q, c) := \sum_t \mathrm{TF}(t, d) \cdot \mathrm{TF}(t, q) \cdot \mathrm{IDF}(t, c)$$

The intuition is to reward terms that occur frequently in document and query ($\mathrm{TF}(t, d)$ and $\mathrm{TF}(t, q)$ high) and are rare among all documents ($\mathrm{IDF}(t, c)$ high for rare terms). The question is: what is the probabilistic root of this intuitive and widely used scoring function?

For answering this question, we summarize the discussion presented in [52]. We start with a basic decomposition of the document probability $P(d \mid c)$:

$$P(d|c) = \prod_{t \text{ IN } d} P(t|c) = \prod_{t \in d} P(t|c)^{n(t,d)} = \prod_{t \in d} P(t|c)^{\mathrm{tf}_d} \quad (2)$$

In the *set-based* product, the exponent of the single event probability is the number of term occurrences (see Equation (1)). Next, we apply the logarithm to the fraction $P(d \mid q)/P(d \mid c)$. The rationale of this fraction is discussed in [24]; the fraction is related to the LM-based approach to IR, and is also justified by divergence-based retrieval [54].

$$\log \frac{P(d \mid q)}{P(d \mid c)} = \log \prod_{t \text{ IN } d} \frac{P(t \mid q)}{P(t \mid c)} = \sum_{t \in d} \log \left(\left(\frac{P(t \mid q)}{P(t \mid c)} \right)^{\mathrm{tf}_d} \right)$$

The next transformation is based on a query-term assumption: for non-query terms, $P(t \mid q) = P(t \mid c)$, whereas for query terms, $P(t \mid q) = 1$. The first setting reflects the assumption usually applied to avoid the ‘zero-probability problem’: for terms that do not occur in the query, the background (collection) model is applied. The second setting can be viewed as maximizing the impact of the foreground (query) model. This leads to the following rank equivalence:

$$\log \frac{P(d \mid q)}{P(d \mid c)} \overset{\text{rank}}{=} \sum_{t \in d \cap q} \mathrm{tf}_d \cdot \log(1/P(t \mid c)) \quad (3)$$

The right-hand side of Equation (3) is the probabilistic root of TF-IDF. However, there is a gap between the root and TF-IDF, and this fuels the view that TF-IDF is heuristic. The two issues in this gap are:

(1) tf_d: The *total count* of term occurrences corresponds to assuming *independence*. The retrieval quality for $\mathrm{TF}_{\text{total}}(t, d) := \mathrm{tf}_d$ is known to be poorer than for $\mathrm{TF}_{\text{BM25}}(t, d) := \mathrm{tf}_d/(\mathrm{tf}_d + K_d)$. Explaining the dependence assumption underlying the BM25-TF is important to close the gap between foundations and heuristics.
(2) $P(t \mid c)$: The *maximum-likelihood* estimate of this probability is based on *counting the occurrences* of term t. The IDF, however, is based on *counting the documents* in which the term occurs (Section 2.5.2).

2.5.4. *BM25: best-match version 25*

The BM25 retrieval model [1] is the main instantiation of the probabilistic retrieval model [55]. It is based on the odds of relevance, $P(r \mid d, q)/(1 - P(r \mid d, q))$, where ‘$r$’ denotes the event ‘relevant’, d denotes a document and q denotes a query. TF-IDF can be viewed as an approximation of BM25 for the case of missing relevance information [51, 56]. The following TF quantification is a main component for achieving good retrieval quality:

$$\mathrm{TF}_{\text{BM25}}(t, d) := \frac{\mathrm{tf}_d}{\mathrm{tf}_d + K_d}$$

The parameter K_d is defined as follows:

$$K_d := k_1 \cdot (b \cdot \mathrm{dl}/\mathrm{avgdl} + (1 - b))$$

There is a range of symbols involved in the BM25-TF:

tf_d	Within-document term frequency (count)
K_d	Parameter to pivotize the TF quantification
dl	Document length
avgdl	Average document length
pivdl	Pivoted document length: pivdl := dl/avgdl
k_1, b	Parameters to adjust the pivotization

K_d captures a document length pivotization. Thus, the rise and saturation of $\mathrm{TF}_{\mathrm{BM25}}$ is faster for short than for long documents.

Overall, though the BM25-TF is motivated by the 2-Poisson model [7], TF-IDF and BM25 are viewed as being heuristic.

2.5.5. *On the BM25-TF and the harmonic sum*

The BM25-TF can be expressed as follows:

$$\mathrm{TF}_{\mathrm{BM25}}(t,d) = \frac{\mathrm{tf}_d/K_d}{\mathrm{tf}_d/K_d + 1} = \frac{\mathrm{TF}_{\mathrm{piv}}(t,d)}{\mathrm{TF}_{\mathrm{piv}}(t,d)+1} \qquad (4)$$

$\mathrm{TF}_{\mathrm{piv}}$ is the pivoted TF. The equation shows how to relate the BM25-TF to the expression $n/(n+1)$, and this factor can be expressed as the harmonic sum of Gaussian sums (Appendix 1):

$$\frac{n}{n+1} = \frac{1}{2}\cdot\left(1 + \frac{1}{1+2} + \cdots + \frac{1}{1+\cdots+n}\right) \qquad (5)$$

Equation (5) led to the core of this paper, namely to apply the harmonic sum for modelling analytically the dependence of subsequent event occurrences.

3. INDEPENDENCE VS HARMONY

Linking Equations (3) and (5) leads to showing analytically the effect of the BM25-TF. For 'naive' TF-IDF (total TF count), the total term frequency count, tf_d, can be expressed as the sum $\mathrm{tf}_d = 1 + 1 + \cdots + 1$.

$$\mathrm{tf}_d \cdot \log(1/P(t\,|\,c)) = (1 + 1 + \cdots + 1)\cdot \log\frac{1}{P(t\,|\,c)}$$

This is in contrast to the smarter BM25-TF-IDF:

$$\begin{aligned}&\frac{\mathrm{tf}_d}{\mathrm{tf}_d+1}\cdot\log\frac{1}{P(t\,|\,c)}\\ &\quad= \frac{1}{2}\cdot\left(1+\frac{1}{1+2}+\cdots+\frac{1}{1+\cdots+\mathrm{tf}_d}\right)\cdot\log\frac{1}{P(t\,|\,c)}\end{aligned}$$

The harmonic sum of Gaussian sums makes explicit that the second-occurrence of the event (term) is considered with $1/3$, and the nth occurrence with $1/(1+\cdots+n)$. The impact of subsequent event occurrences decreases in a harmonic Gaussian way. For independence, each occurrence has the same impact, and two decades of IR research confirm that this is suboptimal. Given the framework of harmony assumption, we can now describe in a precise and analytical form what the currently superior TF quantification means with regard to dependence assumption and probability theory.

4. HARMONY ASSUMPTIONS

We introduce *harmony-based dependence assumptions* to be alternatives to the today's most common assumption, the *independence assumption*. For independence, p_t^n is the sequence probability to observe n occurrence of event t that occurs with probability p_t. In more general, $p_t^{\mathrm{a}(n)}$ is the sequence probability, where $\mathrm{a}(n)$ is the *assumption function*. If the assumption function is based on the harmonic sum, then we refer to the assumption as *harmony assumption*.

Table 1 shows the main harmony assumptions, and Fig. 1 illustrates graphically the effect of harmony: the overlap of event occurrences is proportional to α.

4.1. Natural harmony

If the assumption function is the harmonic sum, then we refer to the dependence as 'natural harmony'.

$$\mathrm{a}_{\text{natural-harmony}}(n) := 1 + \frac{1}{2} + \cdots + \frac{1}{n} \qquad (6)$$

Natural harmony is a parameter-free assumptions. More adjustable (i.e. better for parameter learning) is the generalized harmonic sum with parameter α.

4.2. Alpha-harmony (generalized harmony)

If the assumption function is the generalized harmonic sum, then we refer to the dependence as 'generalized harmony' or 'alpha-harmony'.

$$\mathrm{a}_{\text{generalised-harmony},\alpha}(n) := 1 + \frac{1}{2^\alpha} + \cdots + \frac{1}{n^\alpha} \qquad (7)$$

For $\alpha = 1$, generalized harmony is natural harmony. For $\alpha = 0$, the harmonic sum is $n = 1 + 1 + \cdots + 1$, which

TABLE 1. The main harmony assumptions.

Assumption name	Assumption function $\mathrm{a}(n)$	Description/comment
Natural harmony	$1 + 1/2 + \cdots + 1/n$	Harmonic sum
Alpha-harmony	$1 + \frac{1}{2^\alpha} + \cdots + \frac{1}{n^\alpha}$	Generalized harmonic sum; convergent for $\alpha > 1$
Square-root harmony	$1 + \frac{1}{2^{1/2}} + \cdots + \frac{1}{n^{1/2}}$	$\alpha = 1/2$; divergent
Square harmony	$1 + \frac{1}{2^2} + \cdots + \frac{1}{n^2}$	$\alpha = 2$; convergent: $\pi^2/6 \approx 1.645$
Gaussian harmony	$2 \cdot \frac{n}{n+1} = 1 + \frac{1}{1+2} + \cdots + \frac{1}{1+\cdots+n}$	Explains the BM25-TF $\frac{\mathrm{tf}_d}{\mathrm{tf}_d+\mathrm{pivdl}}$

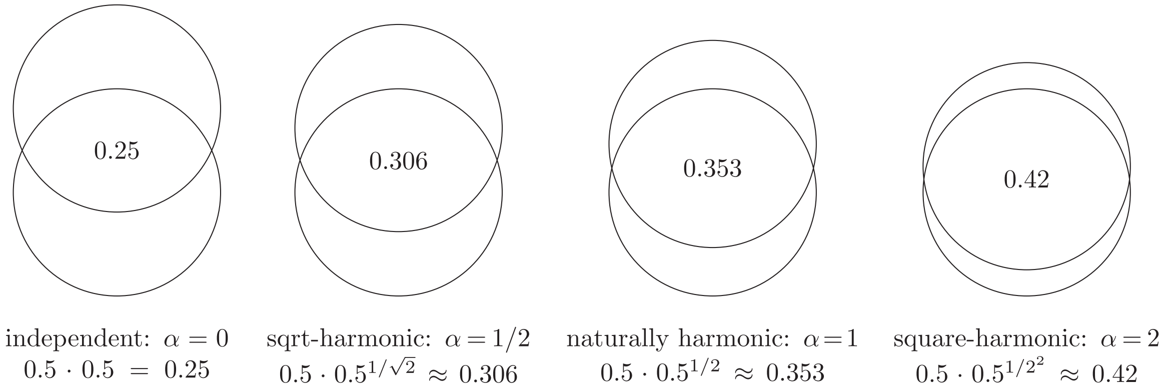

0.25
0.306
0.353
0.42
independent: $\alpha = 0$
$0.5 \cdot 0.5 = 0.25$
sqrt-harmonic: $\alpha = 1/2$
$0.5 \cdot 0.5^{1/\sqrt{2}} \approx 0.306$
naturally harmonic: $\alpha = 1$
$0.5 \cdot 0.5^{1/2} \approx 0.353$
square-harmonic: $\alpha = 2$
$0.5 \cdot 0.5^{1/2^2} \approx 0.42$

FIGURE 1. Illustration of the Overlap for Independent and Harmonic Event Occurrences. The area of each circle corresponds to the single event probability: $p = 0.5$. The overlap becomes larger for growing α (harmony).

corresponds to *independence*. For $\alpha = -\infty$, we obtain *disjointness*, and for $\alpha = +\infty$, *subsumption*.

In more detail, let $1/k^{\alpha}$ be the components of the harmonic sum. For $\alpha = -\infty$, we obtain $1/k^{\alpha} = \infty$ (for $k > 1$). Therefore, the sequence probability is $p^{\infty} = 0$, which corresponds to *disjointness*. For $\alpha = +\infty$, we obtain $1/k^{\alpha} = 0$ (for $k > 1$). Therefore, the sequence probability is $p^1 = p$, which corresponds to *subsumption*.

Since the parameter α can be tuned to fit a distribution, alpha-harmony is a prime choice. The two values $\alpha = 1/2$ and $\alpha = 2$ are special.

4.3. Square-root harmony: $\alpha = 1/2$

Sqrt-harmony ($\alpha = 1/2$) is between natural harmony ($\alpha = 1$) and zero harmony ($\alpha = 0$, independence). Sqrt-harmony views event occurrences to be more overlapping than the independence assumption does, but the overlap is less than for natural harmony (see illustration in Fig. 1). This mid-point between independence and natural harmony turns out to be focal in the experimental study, where we found that sqrt-harmony is the average assumption that represents the dependence between term occurrences, and between user interactions.

4.4. Square harmony: $\alpha = 2$

Square harmony ($\alpha = 2$) reflects stronger dependence than natural harmony ($\alpha = 1$) does.

This assumption is special since $\alpha = 2$ is the smallest natural number for which the harmonic sum is convergent. (The convergence value, $1.645 = \pi^2/6$, is a famous foundation of number theory; Basel problem.) Also, the expression $1/k^2$ is only by a margin greater than half of the expression $2 \cdot 1/(k \cdot (k+1))$, and this suggests the relationship between square harmony and the assumption that motivated this research: Gaussian harmony.

4.5. Gaussian harmony

The harmonic sum of Gaussian sums is

$$\mathrm{a}_{\text{Gaussian-harmony}}(n) := 1 + \frac{1}{1+2} + \cdots + \frac{1}{1+\cdots+n} \qquad (8)$$

This partial harmonic sum comprises the elements where the Gaussian sums are the denominators. The convergence value is as follows (Appendix 1):

$$\frac{2 \cdot n}{n+1} = 1 + \frac{1}{1+2} + \cdots + \frac{1}{1+\cdots+n} \qquad (9)$$

A more compact formulation takes advantage of the Gaussian summation formula:

$$G(k) := k/2 \cdot (k+1)$$

Then, Gaussian harmony can be expressed as follows:

$$\mathrm{a}_{\text{Gaussian-harmony}}(n) = \frac{1}{G(1)} + \frac{1}{G(2)} + \cdots + \frac{1}{G(n)} \qquad (10)$$

Gaussian harmony is the assumption that corresponds to the notion of semi-subsumed events [25], which was the first attempt to explain the dependence underlying the BM25-TF.

4.6. On $\log(1 + \mathrm{tf}_d)$ and harmony assumptions

The question remaining is: how is the log-TF $\log(1 + \mathrm{tf}_d)$ related to harmony assumptions?

To answer this question we recall that the logarithm $\log(x)$ is equal to the integral $\int_1^x 1/z \, \mathrm{d}z$. This leads to a series-based explanation of the logarithmic TF. The explanation originates from the integral-based approximation of the harmonic sum:

$$\log(n) \approx \left[\sum_{k=1}^{n} 1/k \right] - (0.5772 + 1/(2 \cdot n))$$

Here, the Euler–Mascheroni constant 0.5772 and the factor $1/(2 \cdot n)$ approximate the difference between the sum of rectangles and the area under $1/z$.

The relationship between the log and the harmonic sum solves two problems. First, it delivers a semantics of $\log(1 + \mathrm{tf}_d)$ in the sense that the log-TF quantification corresponds to the harmonic sum from 1 to $1 + \mathrm{tf}_d$ minus the factor $0.5772 + 0.5/(1 + \mathrm{tf}_d)$. Therefore, we can embed the log-TF into the assumption spectrum using the name *ln-harmony*. Secondly, the logarithm can be utilized to approximate the value of the harmonic sum, and this reduces the computational complexity.

Future research will investigate how to embed and utilize integral-based expressions related to the generalized harmonic sum. Learning from the approach for the natural harmonic sum and the logarithm, approximations can be developed starting with the integral $\int_1^x 1/z^\alpha \, \mathrm{d}z = 1/(1-\alpha) \cdot (x^{1-\alpha} - 1)$. For the purpose of this paper, it shall be sufficient to investigate series-based assumptions plus the log-based quantification. Future research directions include the approximations of series-based assumptions.

4.7. Other dependence assumptions

In principle, any arbitrary function $f()$ in $p^{f(k)}$ could be used to represent a form of dependence. Therefore, the question is why one would want to use a restricted spectrum of functions, restricted to harmonic sums. There is a particular charm when modelling dependence assumptions via series-based functions such as the harmonic sum. They lead to self-explanatory decay models, and the decay is relatively slow (compared with the thin tails of exponential functions). Of course, any function could be considered and compared with the harmonic sums, and research on dependence functions for IR and SN is gaining momentum.

Regarding a spectrum of dependence assumptions, it is even advisable to define a small set of parameter-free assumptions. This is because people need reference points to utilize mathematical frameworks that are open and general; too general means too many options, which means too complicated. Parameter-free models are easier to exchange and to agree on. Therefore, we have laid out a clear terminology for $\alpha \in \{0.5, 1, 2\}$, for the Gaussian harmonic sum and for embedding $\log(1 + \mathrm{tf}_d)$ (ln-harmony). To briefly indicate that there are other important harmonic sums that are not captured by a value of α, we consider the harmonic sums of primes.

$$1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{5} + \frac{1}{7} + \frac{1}{11} + \frac{1}{13} + \cdots + \frac{1}{\mathrm{prime}(n)}$$

This harmonic sum is interesting because it is smaller than the harmonic sum, but it is still divergent, whereas the harmonic sum is convergent for $\alpha < 1$. As we will find in the next section, the border between convergence and divergence is of particular importance.

4.8. From total harmony to total disharmony

It is challenging to achieve a clear and useful terminology regarding dependence assumptions. Table 2 shows a spectrum of selected assumptions from total harmony (subsumption, $\alpha = +\infty$), over zero harmony (independence, $\alpha = 0$), to total disharmony (disjointness, $\alpha = -\infty$). The greater α, the more likely is the co-occurrence of events (the overlap, see Fig. 1).

TABLE 2. Spectrum of assumptions: from total harmony over zero harmony to total disharmony.

α	Assumption name	Exponent
$+\infty$	Total harmony (subsumption)	1
$+2$	Square harmony	$\sum_{k=1}^{n} \frac{1}{k^2}$
	Gaussian harmony: convergent!	$2 \cdot \frac{n}{n+1}$
	ln-harmony: divergent!	$\ln(n+1)$
$+1$	Natural harmony	$\sum_{k=1}^{n} \frac{1}{k}$
$+1/2$	Sqrt-harmony	$\sum_{k=1}^{n} \frac{1}{\sqrt{k}}$
0	Zero harmony (independence)	n
$-1/2$	Sqrt-disharmony	$\sum_{k=1}^{n} \sqrt{k}$
-1	Natural disharmony	$\frac{n}{2} \cdot (n+1)$
-2	Square-disharmony	$\sum_{k=1}^{n} k^2$
$-\infty$	Total disharmony (disjointness)	∞

For independence, the assumption function is $\mathrm{a}(k) = k$. We refer to assumptions where $0 < \alpha \leq 1$ as harmonic, whereas assumptions where $\alpha > 1$ are over-harmonic (and convergent). For $\alpha \leq 1$, the harmonic sum is divergent. The spectrum illustrates that the two standard TF quantifications, the BM25-TF (Gaussian harmony) and the log-TF (ln-harmony), are the two sides of the border between convergent and divergent. For divergence, any of many event occurrences does matter, whereas for convergence, there is an upper ceiling. To achieve a memorisable terminology, there are systematic names for selected alpha values: $\alpha \in \{-\infty, -2, -1, -0.5, 0, 0.5, 1, 2, +\infty\}$.

The next sections discuss the effect of harmony assumptions when computing sequence and frequency probabilities.

5. SEQUENCE PROBABILITIES

Figure 2 illustrates the nature of sequence probabilities.

The curves illustrate that for harmony ($\alpha > 0$), the sequence probabilities are greater than for independence, whereas for disharmony ($\alpha < 0$), they are smaller. The curves are for

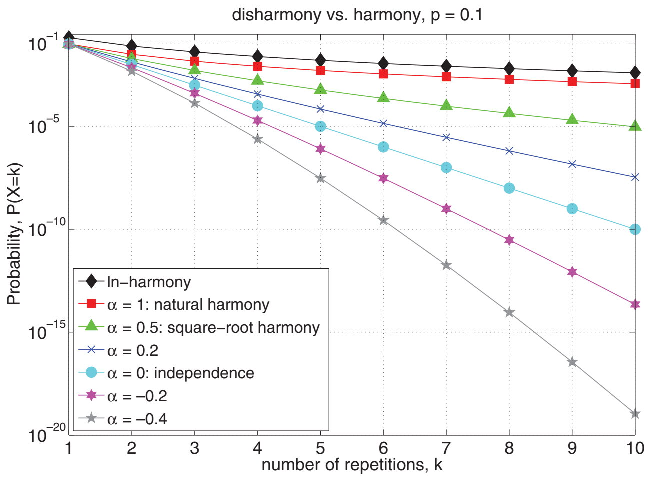


FIGURE 2. Illustration of Sequence Probabilities. $\alpha > 0$: probabilities greater than for independence. $\alpha < 0$: probabilities smaller than for independence.

the event probability $p = 0.1$, and for event frequencies $k \in \{1, \ldots, 10\}$. For independence, the sequence probability is p^k, for example, $p^5 = 10^{-5}$. For sqrt-harmony, it takes $k = 10$ event occurrences for the sequence probability to drop to a value close to 10^{-5}. This is because the harmonic sum of square roots, the exponent is $h_{\alpha=0.5}(10) \approx 5.02$.

Of particular interest for our study are the assumptions between zero and square harmony ($0 < \alpha < 2$). Setting $\alpha > 0$ reflects that word occurrences (and user interactions) occur more frequently together than independence tells. The upper value, $\alpha = 1$, is a strong dependence assumption. The experimental study reports that the middle between independence and natural harmony, namely sqrt-harmony, turns out to be the assumption that models best the average dependence observed in test datasets.

Having explored the effect on sequence probabilities, the next section investigates the effect on frequency probabilities.

6. FREQUENCY PROBABILITIES

It is essential to define and investigate the effect of dependence assumptions on frequency probabilities. There are many potential models, and for the purpose of this paper, we focus the discussion on one of the main models of probability theory: the binomial probability.

6.1. Binomial probability

The binomial probability, and its approximation, the Poisson probability, are the first choice for a probabilistic model of a frequency. For making the case of this paper, it is sufficient to focus on the binomial probability:

$$P_{\mathrm{binomial},n,p}(k) := \binom{n}{k} \cdot p^k \cdot (1-p)^{(n-k)}$$

The case for the Poisson probability is a direction of future research.

6.2. Generalized binomial probability

We base the generalization of the binomial probability on replacing the integer exponents, k and $n - k$, by the assumption functions $\mathrm{a}_1(k)$ and $\mathrm{a}_0(n-k)$. We distinguish between the asymmetric and the symmetric generalization. The asymmetric generalization supports different dependence assumptions for ‘event occurs’ and ‘event does not occur’, whereas the symmetric generalization applies the same assumption, i.e. for the symmetric case, $\mathrm{a} \equiv \mathrm{a}_1 \equiv \mathrm{a}_0$.

The generalization requires to make explicit the normalizing constant Ω, which is usually omitted for independence, since $\Omega = 1$. This follows from the binomial theorem: $1 = (p - (1 - p))^n = \sum_{k=0}^{n} \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$.

Definition 6.1 (Normalizing Constant Ω).

$$\Omega_{n,p,a_1,a_0} := \sum_{k=0}^{n} \binom{n}{k} \cdot p^{a_1(k)} \cdot (1-p)^{a_0(n-k)}$$

Then, the asymmetric generalization is as follows.

Definition 6.2 (Asymmetric Generalized Binomial Probability).

$$P_{n,p,a_1,a_0}(k) := \frac{1}{\Omega} \cdot \binom{n}{k} \cdot p^{a_1(k)} \cdot (1-p)^{a_0(n-k)}$$

Regarding the computation of Ω, for small n, a table approach is sufficient; for large n, approximations can be applied.

The symmetric generalization uses the same assumption function for occurrence and non-occurrence.

Definition 6.3 (Symmetric Generalized Binomial Probability).

$$P_{n,p,a}(k) := \frac{1}{\Omega} \cdot \binom{n}{k} \cdot p^{a(k)} \cdot (1-p)^{a(n-k)}$$

For a being the identity function ($\mathrm{a}(k) = k$), we obtain the traditional binomial probability which assumes independence of event occurrences.

6.3. Harmonic binomial probability

The generalized binomial probability is referred to as harmonic if the assumption function is based on the harmonic sum. Then,

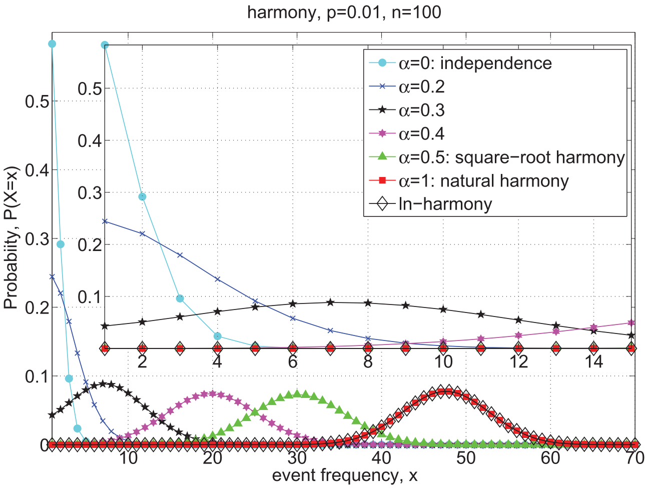


FIGURE 3. Illustration of Frequency Probabilities. Inside graph for $\alpha \in \{0, 0.2, 0.3\}$. The larger the values of α, the more probability mass is shifted to higher frequencies.

the assumption is $\mathrm{a}(k) = h_\alpha(k)$, where $h_\alpha(k)$ is the generalized harmonic sum. For example, for the alpha-harmonic binomial probability function, the assumption function is

$$\mathrm{a}(k) := h_\alpha(k) \qquad \left(h_\alpha(k) = 1 + \frac{1}{2^\alpha} + \cdots + \frac{1}{k^\alpha}\right)$$

In addition to the harmonic assumption functions, we also investigate the logarithmic quantification $\mathrm{a}(k) := \ln(k+1)$.

Figure 3 illustrates alpha-harmonic binomial probabilities (symmetric generalization).

The curves show the values of the probability function $P_{n=100,p=0.01,\alpha}(k)$. The inside graph shows a zoom for $0 \le \alpha \le 0.3$, and the main graph shows selected curves for $0 \le \alpha \le 1$. The curves illustrate that, for harmony, the frequency probabilities of small k are smaller than for independence. Harmony shifts the probability mass to the right, to higher frequencies. For $\alpha = 0.2$, the shift is moderate, and for $\alpha > 0.3$, the curves develop into a bell shape. For total harmony (subsumption, $\alpha = +\infty$), the maximum probability is $P(n/2) = \binom{n}{\lceil n/2 \rceil}/2^n$, because of the binomial theorem and Pascal's triangle.

At the early stage of this research, we investigated asymmetric cases, for example, $a_1(k) := \ln(k+1)$ and $a_0(n,k) := n - \ln(k+1)$. It appears to be easier to find closed forms for asymmetric than for symmetric assumptions. On the other hand, however, asymmetric assumptions cause a misbalance between occurrence and non-occurrence, and this did not lead to useful estimates. Therefore, and because the experimental study was very convincing for the symmetric alpha-harmonic probability, at this stage, we did not engage further with asymmetric variants.

Overall, the definition of the harmonic binomial probability shows that harmony assumptions make the dependendencies between event occurrences explicit in the generalized binomial probability. We conclude this section with a brief look at the multinomial case.

6.3.1. *Multinomial probability*

The multinomial probability is a generalization of the binomial probability. This generalization is with respect to a multi-dimensional space of events. The harmonic assumption functions fits seamlessly into the multinomial probability. Let p_i be the single event probability that event e_i occurs, and the event occurs k_i times. For example, let the sequence $s = e_1, e_3, e_1, e_7, \ldots$ be given, where $k_1 = 2$ for the two occurrences of event e_1, and so forth. The multinomial probability involves the following sequence probability:

$$P(s) = P(e_1, e_3, e_1, e_7, \ldots) = p_1^{k_1} \cdot \ldots \cdot p_M^{k_M}$$

The harmony-based generalization directly applies to the multinomial probability: we apply the assumption functions $\mathrm{a}_i(k_i)$ instead of the total counts k_i.

$$P(e_1, e_3, e_1, e_7, \ldots) = p_1^{\mathrm{a}_1(k_1)} \cdot \ldots \cdot p_M^{\mathrm{a}_M(k_M)}$$

This excursion on the interplay between multinomial probability and harmony concludes the introduction of harmony assumptions, sequence probabilities and frequency probabilities. The next part focuses on the analytical verification and experimental evaluation.

7. ANALYTICAL VERIFICATION

For the analytical verification, we discuss the complexity of harmony, the role of harmony assumptions regarding concepts of IR (TF quantification) and SN's (e.g. centrality), the relationship between harmony and PL, and between harmony and Laplace's law of succession.

7.1. Complexity

Computational complexity is a major concern when capturing the dependence between event occurrences. In general, the computation of the harmonic sum is of linear complexity: $O(n)$. This is not a problem for small n where a tabling approach can be applied. If the computation for larger n were required and computational costs start to be problematic, then approximations of the harmonic sum can be applied.

Another complexity challenge is the computation of the normalization factor Ω. The efficient computation of the normalization constant is a direction of future research (e.g. apply a harmonic variant of the Beta distribution).

7.2. IR models

This paper was motivated by the observation that the BM25-TF, $\mathrm{tf}_d/(\mathrm{tf}_d + K_d)$, can be related to the harmonic sum of Gaussian sums. When exploring the harmonic sum, there was a range of discoveries. First, this led to the analytical description of what TF quantifications capture, namely the dependence between multiple event occurrences expressed via the generalized harmonic sum. This semantics is more general than the semantics presented in [25], where it was reported that the BM25-TF is an assumption between independence and subsumption. This argument is based on the fact that for independence the sequence probability is p^n and for subsumption it is p^1. The harmonic mean of the exponents is $2/(1/1 + 1/n) = 2 \cdot n/(n + 1)$. Because of this property (harmonic mean of independence and subsumption), the assumption has been coined as 'semi-subsumed'. Given the framework of harmony assumptions, semi-subsumption corresponds to Gaussian harmony.

Whereas Gaussian harmony explains the BM25-TF, the experiments brought another assumption into the spotlight: sqrt-harmony. This assumption models best the average dependence in the test datasets. Given this result, the question is whether sqrt-harmony could potentially outperform the BM25-TF? There are indicators pointing at cases where the BM25-TF is suboptimal. The observation is that for long queries, the BM25-TF appears to be suboptimal, i.e. there are other (less saturating) TF quantifications that perform better for long queries. This coincides with our observation that for the IR scenario, there are hardly any terms (and for the SN scenario, there are hardly any users) where $\alpha > 1$. Future publications in the field of IR will report experiments for TF quantifications based on harmony assumptions that are less strong than Gaussian harmony (BM25-TF), and this potentially leads to new standards of TF quantifications.

The overall result from an IR model perspective is that harmony assumptions explain factors that are so far considered as heuristics. Therefore, harmony assumptions close the gap between probability theory, IR models (where $\log(1 + \text{tf}_d)$ and $\text{tf}_d/(\text{tf}_d + K_d)$ are renown TF quantifications), and also classification (where log-tf is applied [12]).

7.3. Social networks

Whereas in IR, the TF quantification compensates for the independence assumption, in SN's, the number of interactions is viewed as a function of previously observed interactions. On the one hand, harmony-based dependence assumptions are a conclusive solution to close the gap between heuristics and sound probability theory, but on the other hand, the potential of applying a harmonic probability theory remains to be explored in future research. For SN's in particular, this concerns tasks such as the link (interaction) prediction problem. It is known that a traditional Poisson process (independence assumption, thin-tail distribution) does not predict interactions appropriately. The hypothesis is that when assuming harmony, it will be possible to devise probabilistic models that perform at least similarly to state-of-the-art approaches based on PL distributions or polya-urn models.

Another interesting relationship between SN's (graphs) and harmony assumptions comes from a notion of 'centrality'. The harmonic distance, i.e. $\sum_{i=1}^{n} 1/i$, where n is the path length between two connected nodes, has been shown to improve the identification of central (popular) nodes [57, 58]. The basic path length can be viewed as assuming independence, whereas the harmonic distance reflects a dependence assumption.

In more general, graph theory can be utilized to model the occurrence of events, and the components of the harmonic sum correspond to the weights associated with the arcs connecting the nodes.

7.4. PL and harmony

The power-law (PL) is a widely known concepts applied for describing a sub-exponential distribution, i.e. a distribution with a tail 'fatter' ('heavier') than the thin tail of the Poisson distribution. It is widely understood and accepted that natural and human-made systems have a fat tail (follow a PL). This holds, for example, for the distribution of

number of cities with population x, and

number of earthquakes with force x.

To be more precise, it is accepted that the distribution is different from Poisson. Or in more general, a thin-tail (exponential) distribution based on an independence assumption does not model the observed distribution. It is also accepted that PL distributions are fat-tail distributions. Whether or not a PL distribution or another distribution best reflects some of the distributions we observe in nature, is an open question. This is where the concept of harmony opens up new pathways to be explored.

Regarding the scenarios IR and SN's, a PL distribution can be observed for the following distributions:

number of documents
in which term t occurs k_t times

number of recipients
with whom user u interacts k_u times

Given that the frequencies follow a PL distribution, we consider PL-based distributions as a candidate model for the experiments. From an analytical point of view very interesting is that the harmonic binomial model brings forward an explicit, self-explanatory way for capturing dependencies, whereas dependencies are not explicit in the PL.

7.5. Laplace's law of succession

The Laplace law of succession is a model to capture that the occurrence of an event affects the probability of the event to occur again. Let m_t be the number of past occurrences of event t, and let M be the number of past trials. Then, given k_t new occurrences in K new trials, the single event probability is

$$P(t) = \frac{m_t + k_t}{M + K}$$

In sequence probabilities, the single event probabilities vary. For example, consider the following probability of a sequence of event occurrences where one event (e.g. one word) occurs twice.

$$P_{\text{Laplace}}(t_1, t_1, t_2) = \frac{m_{t_1}}{M} \cdot \frac{m_{t_1} + 1}{M + 1} \cdot \frac{m_{t_2}}{M}$$

This sequence probability is in contrast to the case of independence:

$$P_{\text{independence}}(t_1, t_1, t_2) = \frac{m_{t_1}}{M} \cdot \frac{m_{t_1}}{M} \cdot \frac{m_{t_2}}{M} = \left(\frac{m_{t_1}}{M}\right)^2 \cdot \frac{m_{t_2}}{M}$$

The Laplace law is also the foundation of polya-urn models that have been applied in the context of SN's. Regarding IR, the Dirichlet compound multinomial distribution [17] has been proposed to mirror the dependence of term occurrences. The concept of harmony can be viewed as an alternative to the Laplace law, where harmony adapts the exponent of the single event probability rather than the event and trial counts.

8. EXPERIMENTAL EVALUATION

In this section, we firstly elicit the duality between IR and SN, and this explains the type of event and frequency probabilities to be studied. Then, we describe the data sets employed. Section 8.3 discusses the distribution of alpha's and some details of the experiments. Section 8.4 summarizes the overall result.

8.1. Duality between IR and SN

For IR, the main event is 'word occurs in document'. For SN, it is 'user interacts with recipient'. The respective probabilities and frequencies are

IR	SN
$p_t = n_t / N_{\text{Words}}$	$p_u = n_u / N_{\text{Interactions}}$
n_t: number of times term t occurs	n_u: number of times user u interacts
N_{Words}: number of trials	$N_{\text{Interactions}}$: number of trials
tf_d: within-document term frequency	uf_r: within-recipient user frequency
dl: document length: number of words in document	rl: recipient 'length': number of interactions of recipient

Given the single event probabilities p_t (p_u), we fit α for each term t (user u). Mathematically, this means that for α, we maximize the value of the log-likelihood ratio test (Appendix 3). In a formal way, let $\hat{\ell}$ be the likelihood function. Then, the optimization is expressed as finding the maximum-likelihood estimates (mle) of α: $\{\hat{\alpha}_{\text{mle}}\} \subseteq \{\arg\max(\alpha)\hat{\ell}(\alpha \mid x_1, \ldots, x_n)\}$. The x_i correspond to the term occurrences in document i (user interactions with recipient i). With regard to binomial and observed probabilities, we search for the α's that best satisfy the following approximations:

$$P_{\text{dl}, p_t, \alpha_t}(k_t) \approx P_{\text{obs}}(\text{number of documents where } \text{tf}_d = k_t)$$
$$P_{\text{rl}, p_u, \alpha_u}(k_u) \approx P_{\text{obs}}(\text{number of recipients where } \text{uf}_\text{r} = k_u)$$

The document length $n = \text{dl}$ corresponds to the number of trials, and p_t is the probability that term t occurs. For SN, the dual formulation is with respect to the recipient 'length' $n = \text{rl}$, and p_u is the probability that user u interacts.

When computing the divergence between model distribution and observed distribution, it is useful to transform to the following setting:

IR: $n = n_t$ and $p_d = 1/N_{\text{Documents}}$
SN: $n = n_u$ and $p_{\text{recipient}} = 1/N_{\text{Recipients}}$

This parameter setting leads to a more efficient computation, since we consider the same single event probability for each term (user, respectively). The setting is justified since the product $n_t \cdot 1/N_{\text{Documents}}$ is equal to $\text{avgdl} \cdot n_t / N_{\text{Words}}$, where $\text{avgdl} = N_{\text{Words}} / N_{\text{Documents}}$.

8.2. Data sets

For the experimental study, we employed two data sets:

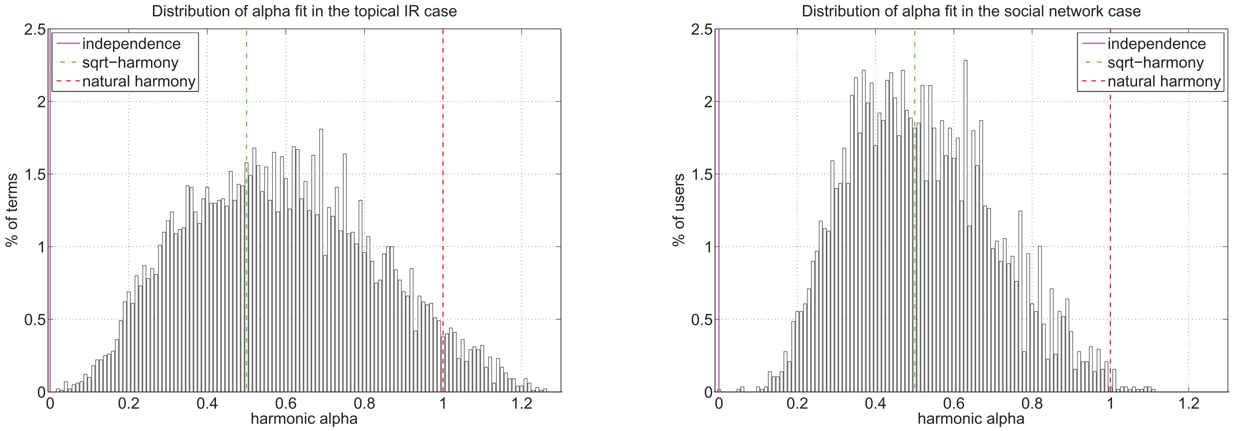
Dataset		
TREC-2	10 000	terms (pruned)
Meneame	5780	users (pruned)

[59] provides details about the Meneame data set and [60] describes the TREC-2 data collection. The terms and users were selected (pruned) to avoid effects from sparse or noisy data. We excluded terms/users that are very rare (too little evidence) or very frequent (users such as owner or administrator who interact with many more users than the ordinary user; terms that occur in most documents). Appendix 2 provides more details.

8.3. Distribution of alpha's

Figure 4 shows graphs and tables for illustrating and discussing the experimental results.

One of the most interesting aspect was to investigate which alpha's generate the best fit between the model (the harmonic binomial probability) and the observed distributions. Figure 4



X vs Y	p-value	X better than Y	no difference	Y better than X
PL vs α-harmony	0.10	21.54%	65.12%	13.34%
	0.05	18.39%	71.80%	9.81%
	0.01	13.54%	80.48%	5.98%
independence vs α-harmony	0.10	0.04%	31.78%	68.18%
	0.05	0.04%	41.66%	58.31%
	0.01	0.03%	57.34%	42.63%
independence vs PL	0.10	0.11%	32.13%	67.76%
	0.05	0.08%	41.78%	58.14%
	0.01	0.03%	57.49%	42.48%
independence vs natural harmony	0.10	67.30%	22.22%	10.48%
	0.05	66.49%	25.73%	7.78%
	0.01	64.74%	31.36%	3.90%
independence vs ln-harmony	0.10	75.40%	17.81%	6.79%
	0.05	74.43%	20.42%	5.16%
	0.01	72.13%	25.22%	2.65%
independence vs sqrt-harmony	0.10	16.75%	32.18%	51.07%
	0.05	16.35%	40.75%	42.90%
	0.01	15.85%	56.12%	28.03%

X vs Y	p-value	X better than Y	no difference	Y better than X
PL vs α-harmony	0.10	7.49%	82.68%	9.83%
	0.05	5.78%	87.85%	6.37%
	0.01	4.13%	92.87%	2.99%
independence vs α-harmony	0.10	0.00%	44.39%	55.61%
	0.05	0.00%	55.67%	44.33%
	0.01	0.00%	71.42%	28.58%
independence vs PL	0.10	0.02%	44.45%	55.54%
	0.05	0.02%	55.78%	44.20%
	0.01	0.00%	71.80%	28.20%
independence vs natural harmony	0.10	78.17%	20.00%	1.83%
	0.05	77.13%	22.06%	0.81%
	0.01	75.42%	24.50%	0.09%
independence vs ln-harmony	0.10	87.01%	12.51%	0.48%
	0.05	85.38%	14.38%	0.24%
	0.01	83.51%	16.47%	0.02%
independence vs sqrt-harmony	0.10	15.62%	46.97%	37.40%
	0.05	15.14%	58.58%	26.28%
	0.01	14.29%	73.98%	11.73%

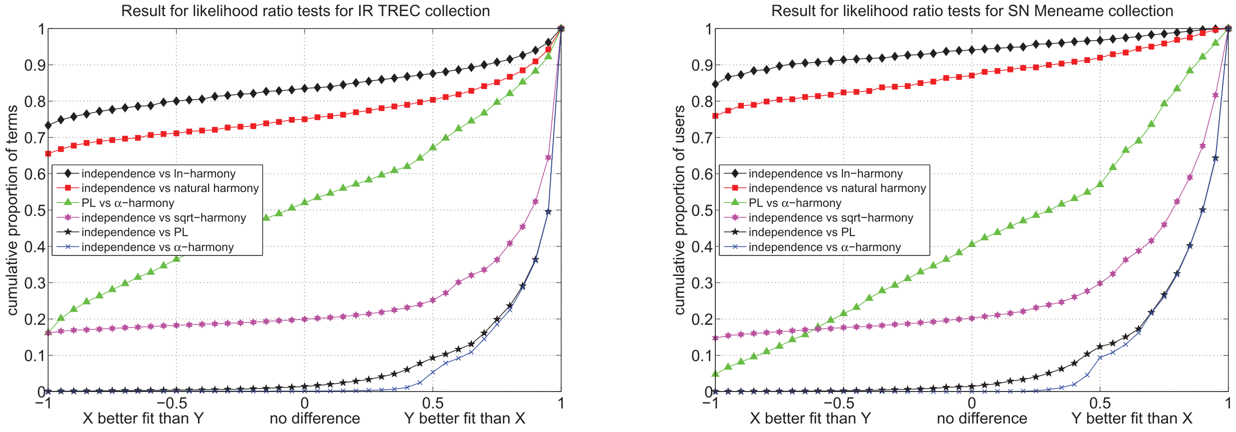


FIGURE 4. IR (left) and SN (right). Distribution of alpha's (top) and log-likelihood ratio (middle and bottom).

(top) shows the distributions of the alpha's. The means and standard variations are

	Mean	stdv
TREC-2	0.58	0.24
Meneame	0.52	0.18

The averages are close to sqrt-harmony ($\alpha = 0.5$), marking it to be the best parameter-free assumption. For many terms (users), sqrt-harmony reflects the underlying dependence between term occurrences (user interactions). For IR, 56.5% of the terms, and for SN, 63.1% of the users, have a dependence that lies in the interval $0.4 \leq \alpha \leq 0.8$.

The tables in the middle of Fig. 4 show the numerical values underlying the graphical illustration at the bottom of the figure. The comparison is based on the log-likelihood ratio test (see Section 8.4 for details). For the comparison, we pair PL vs harmony, independence vs harmony and independence vs PL. Moreover, we pair independence vs the main parameter-free harmony assumptions: natural harmony, ln-harmony and sqrt-harmony.

The comparisons show the proportion of terms and users for which candidate distribution X is better than distribution Y. The notion 'Y=independence' corresponds to applying the traditional, independence-based binomial distribution, whereas '$Y = \alpha$-harmony' refers to the generalized binomial distribution (α learnt for each term and user). For 'Y=sqrt-harmony', the setting $\alpha = 0.5$ is applied for all terms (users).

The most remarkable result is the one for 'independence vs sqrt-harmony' (pink line in coloured graphs). For example, for P-value = 0.1, sqrt-harmony is better for 51.07% of the terms, whereas independence is better for only 16.75%.

8.4. Conclusions and overall result

The main conclusions are:

(1) For both scenarios, IR and SN, assuming harmony is more appropriate and versatile than assuming independence.

(2) Alpha-harmony outperforms independence; this is as expected, since there is the tuning parameter α, whereas independence is parameter-free. Therefore, we also report the results for parameter-free harmony assumption, and compare the PL (parameter α) vs harmony.

(3) Sqrt-harmony outperforms independence; this is evident from the distribution of alpha's (left and right plots at top of Fig. 4): the average is $\alpha \approx 0.5$.

(4) Natural harmony is outperformed by independence. This shows that natural harmony is a too strong dependence assumption, at least for the data sets considered. The distribution of alpha's shows that there are only few terms and users for which such a strong dependence holds.

(5) The harmonic binomial probability shows about the same performance as the PL-based probability.

(6) The majority of observed dependencies lies between independence, $\alpha = 0$, and natural harmony, $\alpha = 1$. For IR, 56.5% of the terms, and for SN, 63.1% of the users, have a dependence that lies in the interval $0.4 \leq \alpha \leq 0.8$.

(7) The experiments indicate that there is an underlying law of harmony in both, IR (term frequencies) and SN (user frequencies).

The overall result is that harmony assumptions are an appropriate and relatively intuitive framework to model dependencies between event occurrences.

9. SUMMARY AND OUTLOOK

This paper proposes *harmony-based dependence assumptions* to model the dependence between the occurrences of an event. Harmony assumptions help to analytically describe the dependence assumption inherently modelled by 'heuristic' parameters (whether or not the parameters are heuristic is usually a major discussion) in otherwise probabilistic models. We coin the notion *harmony* because of the usage of the harmonic sum as the decaying exponent in a sequence probability. Whereas for independence, the sequence probability is $p^{1+1+\cdots+1} = p^n$, for natural harmony, it is $p^{1+1/2+\cdots+1/n}$.

The main contributions of this paper are the groundwork and definition of the main harmony assumptions, and the experimental study of dependencies as they occur in IR and SN. There is a clear terminology to refer to selected assumptions on the harmony scale (Table 2). In order of increasing dependence, the assumptions are

zero harmony (independence, $\alpha = 0$) $\prec$
sqrt-harmony ($\alpha = 1/2$) $\prec$
natural harmony ($\alpha = 1$) $\prec$
ln-harmony (divergent) $\prec$
Gaussian harmony (convergent) $\prec$
square harmony ($\alpha = 2$) $\prec$
total harmony (subsumption, $\alpha = +\infty$)

There are few, well-defined and parameter-free assumptions placed between the traditional assumptions independence and subsumption. Notably, the two TF quantifications log-TF (ln-harmony) and BM25-TF (Gaussian harmony) frame the border between divergent and convergent.

The concept of harmony meets the general perception that event occurrences are bursty. This is the subjective feeling many people articulate when waiting for an event to occur: while there are many periods where the event does not occur, suddenly, there is a period in which the event occurs several times. The probability that the event occurrences are close to each other is greater than the traditional, independence-based binomial probability tells.

For IR, term frequencies (term occurrences in a document) greater than one are more likely than if the term occurrences were independent. The same pattern can be observed for SN, where user frequencies (user interactions with a recipient) greater than one are more likely than for independent interactions. Given the framework of harmony assumptions, we can capture the dependence as alpha values. A single value, namely the average alpha value, reflects the average dependence. This is a key characteristics of data that can be used alongside averages, deviations and other values.

A particularly interesting dependence assumption is *square-root harmony* ($\alpha = 1/2$). It lies between independence ($\alpha = 0$) and natural harmony ($\alpha = 1$). For the experiments, this assumption models best the average dependence.

Another interesting assumption is *Gaussian harmony*, $2 \cdot n/(n+1)$. It delivers a series-based explanation of the BM25-TF, and this closes the long-standing gap between probability theory and this TF quantification.

Overall, the framework of harmony assumptions supports the idea to choose an appropriate dependence assumption instead of the independence assumption and heuristics. While making harmony assumptions is less complex than modelling with Markov chains or Bayesian networks, it is a complementary concept, and future work will combine it with models that are tailored to capturing explicitly the dependencies between *different events*.

On the theoretical side, harmony assumptions fit seamlessly into existing probability theory, leading to a theory that is more capable in capturing the dependence between event occurrences in large-scale applications. On the pragmatic side, harmony assumptions seem to fit into the real world, explaining effects we sometimes feel but find difficult to explain, namely that a rare event suddenly occurs several times.

Returning to the boulders at the entrance of a cave, we still do not know whether the boulder falls today or another day. But, what we can model is that if one boulder falls, then, for harmony, it is more likely that another boulder falls as well, and for disharmony, it is less likely, i.e. the other boulders will hold still. It is currently hypothetical, but it is potentially possible, that for natural and human-made systems, characteristic alpha values can be found. This will help to formulate general laws that can guide the modelling of systems.

ACKNOWLEDGEMENTS

We would like to acknowledge Aris Gionis and Paolo Boldi. Aris' feedback sharpened the terminology, and also forecasts the modelling of harmony in graph theory. This coincides with the way Paolo utilizes the harmonic sum in networks to compute a centrality score of nodes. This work was mainly developed while T.R. was a visiting researcher at Yahoo Labs and Barcelona Media. We would like to acknowledge the CJ reviewers who provided valuable feedback and recommendations.

FUNDING

This work was funded by Yahoo Labs and Barcelona Media.

REFERENCES

[1] Robertson, S., Hancock-Beaulieu, M. and Gatford, M. (1994) Okapi at TREC-3. *Text REtrieval Conference*, Gaithersburg, Maryland, USA.

[2] Robertson, S. E., Walker, S. and Hancock-Beaulieu, M. (1995) Large test collection experiments on an operational interactive system: Okapi at TREC. *Inf. Process. Manage.*, **31**, 345–360.

[3] Penrose, R. (1994) *Shadows of the Mind—A Search for the Missing Science of Consciousness*. Oxford University Press.

[4] Margulis, E. (1992) N-Poisson Document Modelling. In Belkin, N., Ingwersen, P. and Pejtersen, M. (eds), *Proc. 15th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, Copenhagen, Denmark, pp. 177–189. ACM, New York.

[5] Church, K. and Gale, W. (1995) Poisson mixture. *Nat. Lang. Eng.*, **1**, 163–190.

[6] Church, K. and Gale, W. (1995) Inverse Document Frequency (idf): A Measure of Deviation from Poisson. *Proc. 3rd Workshop on Very Large Corpora*, pp. 121–130.

[7] Robertson, S. E. and Walker, S. (1994) Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval. In Croft, W.B. and van Rijsbergen, C.J. (eds), *Proc. 17th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, London, pp. 232–241. Springer.

[8] Singhal, A., Buckley, C. and Mitra, M. (1996) Pivoted Document Length Normalisation. In Frei, H., Harmann, D., Schäuble, P. and Wilkinson, R. (eds), *Proc. 19th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, New York, pp. 21–39. ACM.

[9] Lewis, D. D. (1998) Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval. *ECML'98: Proc. 10th European Conf. on Machine Learning*, London, UK, pp. 4–15. Springer.

[10] Yang, Y. (1999) An evaluation of statistical approaches to text categorization. *Inf. Retr.*, **1**, 69–90.

[11] Sebastiani, F. (2002) Machine learning in automated text categorization. *ACM Comput. Surv.*, **34**, 1–47.

[12] Lewis, D. D., Yang, Y., Rose, T. and Li, F. (2004) RCV1: a new benchmark collection for text categorization research. *J. Mach. Learn. Res.*, **5**, 361–397.

[13] Amati, G. and Rijsbergen, C. J. (2002) Term Frequency Normalization via Pareto Distributions. In Crestani, F., Girolami, M. and Rijsbergen, C.J. (eds), *24th BCS-IRSG European Colloquium on IR Research*, Glasgow, Scotland.

[14] Amati, G. and van Rijsbergen, C. J. (2002) Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*, **20**, 357–389.

[15] He, B. and Ounis, I. (2005) A Study of the Dirichlet Priors for Term Frequency Normalisation. *Proc. 28th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, New York, NY, USA, pp. 465–471. ACM Press.

[16] He, B. and Ounis, I. (2005) Term Frequency Normalisation Tuning for BM25 and DFR Models. *ECIR*, Santiago de Compostela, Spain, pp. 200–214.

[17] Xu, Z. and Akella, R. (2008) A New Probabilistic Retrieval Model Based on the Dirichlet Compound Multinomial Distribution. In Myaeng, S.-H., Oard, D.W., Sebastiani, F., Chua, T.-S. and Leong, M.-K. (eds), *SIGIR*, pp. 427–434. ACM.

[18] Hou, Y., He, L., Zhao, X. and Song, D. (2011) Pure High-Order Word Dependence Mining via Information Geometry. *Advances in Information Retrieval Theory: 3rd Int. Conf., Proc., ICTIR*, 2011, Bertinoro, Italy, September 12–14, pp. 64–76. Springer, New York Inc.

[19] Ponte, J. and Croft, W. (1998) A Language Modeling Approach to information Retrieval. In Croft, W.B., Moffat, A., van Rijsbergen, C.J., Wilkinson, R. and Zobel, J. (eds), *Proc. 21st Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, New York, pp. 275–281. ACM.

[20] Hiemstra, D. (2000) A probabilistic justification for using tf.idf term weighting in information retrieval. *Int. J. Digit. Libr.*, **3**, 131–139.

[21] Bruza, P. and Song, D. (2003) A Comparison of Various Approaches for Using Probabilistic Dependencies in Language Modeling. *SIGIR*, pp. 419–420. ACM.

[22] Robertson, S. (2005) On event spaces and probabilistic models in information retrieval. *Inf. Retr. J.*, **8**, 319–329.

[23] Roelleke, T. and Wang, J. (2006) A Parallel Derivation of Probabilistic Information Retrieval Models. *ACM SIGIR*, Seattle, USA, pp. 107–114.

[24] Roelleke, T. and Wang, J. (2008) TF-IDF Uncovered: A Study of Theories and Probabilities. *ACM SIGIR*, Singapore, July, pp. 435–442.

[25] Wu, H. and Roelleke, T. (2009) Semi-subsumed Events: A Probabilistic Semantics for the BM25 Term Frequency Quantification. *ICTIR (International Conference on Theory in Information Retrieval)*. Springer.

[26] Fuhr, N. and Roelleke, T. (1997) A probabilistic relational algebra for the integration of information retrieval and database systems. *ACM Trans. Inf. Syst.*, **14**, 32–66.

[27] Benjelloun, O., Sarma, A. D., Halevy, A. Y. and Widom, J. (2006) ULDBs: Databases with Uncertainty and Lineage. *VLDB*, Seoul, Korea, pp. 953–964.

[28] Roelleke, T. and Fuhr, N. (1998) Information Retrieval with Probabilistic Datalog. In Crestani, F., Lalmas, M. and Rijsbergen, C.J. (eds), *Uncertainty and Logics—Advanced Models for the Representation and Retrieval of Information*. Kluwer Academic Publishers.

[29] Fuhr, N. and Roelleke, T. (1998) HySpirit—A Probabilistic Inference Engine for Hypermedia Retrieval in Large Databases. In Schek, H.-J., Saltor, F., Ramos, I. and Alonso, G. (eds), *Proc. 6th Int. Conf. on Extending Database Technology (EDBT)*, New York, NY, USA, pp. 24–38. ACM.

[30] Antova, L., Koch, C. and Olteanu, D. (2007) From Complete to Incomplete Information and Back. *SIGMOD Conf.*, Beijing, China, pp. 713–724.

[31] Antova, L., Koch, C. and Olteanu, D. (2007) World-Set Decompositions: Expressiveness and Efficient Algorithms. *ICDT*, Barcelona, Spain, pp. 194–208.

[32] Dalvi, N. N., Miklau, G. and Suciu, D. (2005) Asymptotic Conditional Probabilities for Conjunctive Queries. *ICDT*, Edinburgh, Scotland, pp. 289–305.

[33] Dalvi, N. N. and Suciu, D. (2007) Efficient query evaluation on probabilistic databases. *VLDB J.*, **16**, 523–544.

[34] Theobald, M., Weikum, G. and Schenkel, R. (2004) Top-k Query Evaluation with Probabilistic Guarantees. *VLDB*, Toronto, Ontario, Canada, pp. 648–659.

[35] Barabâsi, A., Jeong, H., Néda, Z., Ravasz, E., Schubert, A. and Vicsek, T. (2002) Evolution of the social network of scientific collaborations. *Phys. A: Statist. Mech. Appl.*, **311**, 590–614.

[36] Newman, M. (2005) Power laws, Pareto distributions and Zipf's law. *Contemp. Phys.*, **46**, 323–351.

[37] Liben-Nowell, D. and Kleinberg, J. (2007) The link-prediction problem for social networks. *J. Am. Soc. Inf. Sci. Technol.*, **58**, 1019–1031.

[38] Lü, L. and Zhou, T. (2011) Link prediction in complex networks: a survey. *Phys. A: Statist. Mech. Appl.*, **390**, 1150–1170.

[39] Murata, T. and Moriyasu, S. (2007) Link Prediction of Social Networks Based on Weighted Proximity Measures. *Web Intelligence, IEEE/WIC/ACM Int. Conf. on*, pp. 85–88. IEEE.

[40] Lü, L. and Zhou, T. (2010) Link prediction in weighted networks: The role of weak ties. *Europhys. Lett.*, **89**, 18001.

[41] Yager, R. (1979) A note on probabilities of fuzzy events. *Inf. Sci.*, **18**, 113–129.

[42] Dubois, D. and Prade, H. (1988) An Introduction to Possibilistic and Fuzzy Logics. In Smets, P., Mamdani, A., Dubois, D. and Prade, H. (eds), *Non-Standard Logics for Automated Reasoning*, pp. 287–326. Academic Press, London.

[43] Pednault, E., Zucker, S. and Muresan, L. (1981) On the independence assumption underlying subjective Bayesian updating. *Artif. Intell.*, **16**, 213–222.

[44] Yu, C., Buckley, C., Lam, K. and Salton, G. (1983) A generalized term dependence model in information retrieval. *Inf. Technol.: Res. Dev.*, **2**, 129–154.

[45] Salton, G., Buckley, C. and Yu, C. (1983) An Evaluation of Term Dependence Models in Information Retrieval. In Salton, G. and Schneider, H.-J. (eds), *Research and Development in Information Retrieval*, Berlin, pp. 151–173. Springer.

[46] Glymour, C. (1985) Independence assumptions and Bayesian updating. *Artif. Intell.*, **25**, 95–99.

[47] Johnson, R. (1986) Independence and Bayesian Updating Methods. In Kanal, L. and Lemmer, J. (eds), *Uncertainty in Artificial Intelligence*, pp. 197–201. North-Holland, Amsterdam.

[48] Croft, W. B. (1986) Boolean queries and term dependencies in probabilistic retrieval models. *J. Am. Soc. Inf. Sci.*, **37**, 71–77.

[49] Grossman, D. A. and Frieder, O. (2004) *Information Retrieval. Algorithms and Heuristics* (2nd edn). The Information Retrieval Series, Vol. 15. Springer.

[50] Baeza-Yates, R. A. and Ribeiro-Neto, B. A. (2011) *Modern Information Retrieval—the Concepts and Technology Behind Search* (2nd edn). Pearson Education Ltd., Harlow, England.

[51] Robertson, S. (2004) Understanding inverse document frequency: on theoretical arguments for idf. *J. Doc.*, **60**, 503–520.

[52] Roelleke, T. (2013) *Information Retrieval Models: Foundations and Relationships*. Synthesis Lectures on Information Concepts, Retrieval, and Services. Morgan & Claypool Publishers.

[53] Luk, R. W. P. (2008) On event space and rank equivalence between probabilistic retrieval models. *Inf. Retri. J.*, **11**, 539–561.

[54] Zhai, C. (2009) *Statistical Language Models for Information Retrieval*. Morgan & Claypool Publishers.

[55] Robertson, S. and Sparck-Jones, K. (1976) Relevance weighting of search terms. *J. Am. Soc. Inf. Sci.*, **27**, 129–146.

[56] Croft, W. and Harper, D. (1979) Using probabilistic models of document retrieval without relevance information. *J. Doc.*, **35**, 285–295.

[57] Boldi, P., Rosa, M. and Vigna, S. (2011) Hyperanf: Approximating the Neighbourhood Function of Very Large Graphs on a Budget. *Proc. 20th Int. Conf. on World Wide Web*, pp. 625–634. ACM.

[58] Boldi, P. and Vigna, S. (2012) Four Degrees of Separation, Really. *Int. Conf. on Advances in Social Networks Analysis and Mining, ASONAM 2012*, Istanbul, Turkey, August 26–29, pp. 1222–1227. IEEE Computer Society.

[59] Kaltenbrunner, A., Gonzalez, G., Ruiz De Querol, R. and Volkovich, Y. (2011) Comparative analysis of articulated and behavioural social networks in a social news sharing website. *New Rev. Hypermed. Multimed.*, **17**, 243–266.

[60] Harman, D. (1995) Overview of the second text retrieval conference (trec-2). *Inf. Process. Manage.*, **31**, 271–290.

[61] Clauset, A., Shalizi, C. and Newman, M. (2009) Power-law distributions in empirical data. *SIAM Rev.*, **51**, 661–703.

APPENDIX 1. GAUSSIAN HARMONY

This section contains the formal proof regarding the relationship between the BM25-TF, $2 \cdot \mathrm{tf}_d/(\mathrm{tf}_d + K_d)$, and the harmonic sum of Gaussian sums.

We show the proof for the case $K_d = 1$, i.e. for a document of average length.

Theorem A.1 (BM25-TF is Gaussian Harmony). *The BM25-TF for the case of a document of average length is equal to the harmonic sum of Gaussian sums.*

$$\frac{2 \cdot n}{n+1} = 1 + \frac{1}{1+2} + \cdots + \frac{1}{1+2+\cdots+n} \tag{A.1}$$

Proof. The proof is via induction. For the proof, we will apply the common Gaussian summation formula:

$$G(n) = 1 + 2 + \cdots + n = n/2 \cdot (n+1)$$

Induction start, $n = 1$:

$$\frac{2 \cdot 1}{1+1} = 1$$

Induction assumption:

$$\frac{2 \cdot n}{n+1} = 1 + \frac{1}{1+2} + \cdots + \frac{1}{1+2+\cdots+n}$$

Induction step, $n \to (n+1)$:

$$\frac{2 \cdot (n+1)}{(n+1)+1} = 1 + \frac{1}{1+2} + \cdots + \frac{1}{1+2+\cdots+(n+1)}$$

Insertion of induction assumption for n:

$$\frac{2 \cdot (n+1)}{(n+1)+1} = \frac{2 \cdot n}{n+1} + \frac{1}{1+2+\cdots+(n+1)}$$

Insertion of Gaussian summation formula: $G(n+1) = (n+1)/2 \cdot (n+2)$:

$$\frac{2 \cdot (n+1)}{(n+1)+1} = \frac{2 \cdot n}{n+1} + \frac{1}{(n+1)/2 \cdot (n+2)}$$

Bring to common denominator:

$$\frac{2 \cdot (n+1) \cdot (n+1)}{(n+2) \cdot (n+1)} = \frac{2 \cdot n \cdot (n+2)}{(n+1) \cdot (n+2)} + \frac{2}{(n+1) \cdot (n+2)}$$

It remains to show the equality of the numerators.

$$2 \cdot (n+1) \cdot (n+1) = 2 \cdot n \cdot (n+2) + 2$$

The following rewriting shows the equality:

$$2 \cdot (n+1)^2 = 2 \cdot (n^2 + 2 \cdot n + 1)$$ □

APPENDIX 2. TERM OCCURRENCES

Table A1 shows a snapshot of the term statistics used for the experimental study. For each term, there are four rows. The first row, labelled with the term (stemming applied), shows the number of documents in which the term occurs k times. The second row (labelled tf) shows the number of term occurrences (product of k and number of documents in which the term occurs). For example, the term 'africa' (third term), occurs once ($k = 1$) in 4584 documents, twice ($k = 2$) in 1462 documents, etc. The probability $P_{\mathrm{obs}}(1) = 0.0062$ is smaller than the Poisson probability $P_{\mathrm{Poisson},\lambda}(1) = 0.0258$. For $k \geq 2$, however, the observed probability is greater than the Poisson probability tells.

The most right column contains the total number of documents in which the term occurs (for africa, 8533), the total number of occurrences (for africa, 19 681) and the average occurrence (the parameter of the Poisson probability). The average is $\lambda(t, c) = n_{\mathrm{Locations}}(t, c)/N_D(c)$, where $N_D(c) = 742\,611$ is the number of *Documents* in collection $c \equiv$ TREC-2. Thus, for africa, $\lambda = 19\,681/742\,611 = 0.0265$.

The pattern of observed and Poisson probabilities illustrates that for $k > 2$ the observed probabilities are greater than the Poisson probability tells.

The pruning selects terms (users) that occur in at least 20 documents (that interact with at least 20 recipients), and the term occurs in at least one document more than once (the user interacts with at least one recipient more than once).

TABLE A1. Term statistics: test collection TREC-2.

k	0	1	2	3	4	5	6	7	8	9	10	
act	628 334	64 628	22 914	9963	5534	3500	2640	1840	1405	1061	792	114 277
tf		64 628	45 828	29 889	22 136	17 500	15 840	12 880	11 240	9549	7920	237 410
p_{obs}	0.846	0.0870	0.0309	0.0134	0.0075	0.0047	0.0036	0.0025	0.0019	0.0014	0.0011	
Poisson	0.7264	0.2322	0.0371	0.0040	0.0003	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	0.3197
antimis	742 516	86	7	1	1	0	0	0	0	0	0	95
tf		86	14	3	4	0	0	0	0	0	0	107
p_{obs}	1.000	0.0001	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	0.0	0.0	0.0	0.0	0.0	0.0	
Poisson	0.9999	0.0001	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	0.0001
africa	734 078	4584	1462	809	550	345	271	182	137	105	88	8533
tf		4584	2924	2427	2200	1725	1626	1274	1096	945	880	19 681
p_{obs}	0.989	0.0062	0.0020	0.0011	0.0007	0.0005	0.0004	0.0002	0.0002	0.0001	0.0001	
Poisson	0.9738	0.0258	0.0003	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	0.0265
control	630 781	69 990	21 259	8836	4444	2747	1625	1113	795	550	471	111 830
tf		69 990	42 518	26 508	17 776	13 735	9750	7791	6360	4950	4710	204 088
p_{obs}	0.849	0.0942	0.0286	0.0119	0.0060	0.0037	0.0022	0.0015	0.0011	0.0007	0.0006	
Poisson	0.7597	0.2088	0.0287	0.0026	0.0002	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	0.2748
compan	506 578	103 554	52 147	30 077	17 788	11 254	7519	5204	3631	2793	2066	236 033
tf		103 554	104 294	90 231	71 152	56 270	45 114	36 428	29 048	25 137	20 660	581 888
p_{obs}	0.682	0.1394	0.0702	0.0405	0.0240	0.0152	0.0101	0.0070	0.0049	0.0038	0.0028	
Poisson	0.4568	0.3579	0.1402	0.0366	0.0072	0.0011	0.0001	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	0.7836
govern	615 963	60 675	26 725	15 013	8899	5579	3619	2492	1666	1163	817	126 648
tf		60 675	53 450	45 039	35 596	27 895	21 714	17 444	13 328	10 467	8170	293 778
p_{obs}	0.829	0.0817	0.0360	0.0202	0.0120	0.0075	0.0049	0.0034	0.0022	0.0016	0.0011	
Poisson	0.6733	0.2663	0.0527	0.0069	0.0007	0.0001	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	0.3956
human	710 486	21 608	5664	2145	1039	639	390	219	197	129	95	32 125
tf		21 608	11 328	6435	4156	3195	2340	1533	1576	1161	950	54 282
p_{obs}	0.957	0.0291	0.0076	0.0029	0.0014	0.0009	0.0005	0.0003	0.0003	0.0002	0.0001	
Poisson	0.9295	0.0679	0.0025	0.0001	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	0.0731
islam	739 052	2108	665	324	188	103	73	35	23	19	21	3559
tf		2108	1330	972	752	515	438	245	184	171	210	6925
p_{obs}	0.995	0.0028	0.0009	0.0004	0.0003	0.0001	0.0001	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	
Poisson	0.9907	0.0092	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	0.0093
medicin	734 560	6117	1012	510	191	97	41	24	31	12	16	8051
tf		6117	2024	1530	764	485	246	168	248	108	160	11 850
p_{obs}	0.989	0.0082	0.0014	0.0007	0.0003	0.0001	0.0001	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	
Poisson	0.9842	0.0157	0.0001	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	0.0160
spy	740 614	1309	320	137	107	54	32	17	12	4	5	1997
tf		1309	640	411	428	270	192	119	96	36	50	3551
p_{obs}	0.997	0.0018	0.0004	0.0002	0.0001	0.0001	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	
Poisson	0.9952	0.0048	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	$<10^{-4}$	0.0048

APPENDIX 3. LIKELIHOOD RATIO TEST

For comparing two distributions regarding their fit to a given dataset, we use the likelihood ratio test as described in [61]. The idea behind this test is to analyse the differences (point-wise) between the log-likelihoods of the data-points within the two distributions (the sum over $k = 1, \ldots, 10$ of $\log P_{\mathrm{X}}(k)/P_{\mathrm{Y}}(k)$). The test hypothesis is that the sum of these differences is close to zero. If the probability (the P-value) of observing the actual difference is small, then one can assume that the sign of the differences indicates which one of the two distributions is a better fit of the data. The tables in Fig. 4 indicate for given P-values the number of cases where a distribution X is better than Y.