Query Expansion with Random Embeddings

Lucas Bernardi

January 2014

Query Expansion

Problem Statement

- In the context of Information Retrieval, Query Expansion is a method to increase recall and/or precision by modifying a user query
- A specific case is the expansion of each term to related terms, increasing recall (and precision as a side effect)
- Example: User query: soccer, expanded query: soccer football maradona
- The Problem: Learn to expand terms

A General Approach

The distributional hypothesis

Words with similar distributional properties have similar meanings.

The geometric metaphor of meaning

Meanings are locations in a semantic space, and semantic similarity is proximity between the locations.

The Vector Space Model Approach

A solution

- An algebraic model for information retrieval and NLP
- Defines a vector space to represent terms
- Defines a dimension for each unique term
- Each term is represented by a semantic vector
- Each component of a semantic vector is the weight of the represented term in the direction of the dimension term
- Weights are a design decision, TF-IDF is widely used
- Compute the related terms of a given term as its k-nearest neighbors
- The vector distance/similarity measure is a design decision, cosine similarity is widely adopted

The Vector Space Model Approach

Some drawbacks

- Curse of dimensionality
- Sparse vectors
- Dynamic vector space

Alternatives

- Use documents as dimensions: Less dimensions, but still sparse vectors and still a dynamic space
- Use topics as dimensions: Less dimensions, dense vectors and fixed space, but how can we define topics and assign weights? Latent Semantic Analysis

The Random Projections Approach

Overcoming drawbacks

- Use a random low dimensional space
- Less dimensions, dense vectors, fixed space
- No semantic assigned to dimensions
- But, how are weights assigned?

The Random Projections Approach

A simple algorithm

- Assign a random k-dimensional vector to each term (index vector)
- Allocate a null k-dimensional vector to each term (semantic vector)
- For each sentence in the corpus compute bigrams (*left right*)
- For each bigram
  - $semantic(left)$ += $index(right)$
  - $semantic(right)$ += $index(left)$

The Random Projections Approach

A simple algorithm: Illustration

The Random Projections Approach

A simple algorithm: a formal description

Given a text documents corpus D we define
$\mathbb{T} = \{\text{All terms in } D\}$, $t = |\mathbb{T}|$
$\mathbb{C} = \{\text{All contexts in } D\}$, where a context is a set of terms
t random k-dimensional vectors $r_i, 1 \leq i \leq t$, $\mathbb{R} = \{r_i\}$
t semantic k-dimensional vectors $s_i, 1 \leq i \leq t$, $\mathbb{S} = \{s_i\}$
A mapping $S : w \in \mathbb{T} \rightarrow \mathbb{S}$, which maps a term to its semantic vector
A mapping $R : w \in \mathbb{T} \rightarrow \mathbb{R}$, which maps a term to its random vector
A mapping $C : w \in \mathbb{T} \rightarrow \{\mathbb{C}\}$, which maps a term to a set containing all the contexts it appears
Then we can express the semantic vector assigned to the term w as:

$$s_w = \sum_{c \in C(w)} \sum_{p \in c} R(p) \tag{1}$$

The Random Projections Approach

Algebraic Interpretation

Defining a co-occurrence matrix $M \in \mathbb{Z}^{t \times t}$, $M_{ij} = \sum_{c \in C(t_i)} \mathbb{1}_c(t_j)$

Each cell i, j in M counts the amount of contexts in D containing the ith and jth terms

Introducing a random matrix $R \in \mathbb{Z}^{t \times k}$, $R_i = R(t_i)$, that is, all random vectors as rows, we can express a semantic vector as:

$$s_i = \sum_{j=1}^{t} R(t_j) M_{ij} = \sum_{j=1}^{t} R_j M_{ij} \tag{2}$$

If we also define the semantic matrix $S \in \mathbb{Z}^{t \times k}$ with $S_i = s_i$, that is, all semantic vectors as rows, we can finally write:

$$S = MR \tag{3}$$

The Random Projections Approach

Algebraic Interpretation

$$S = MR \tag{4}$$

- This equation is a mathematical expression of our initial algorithm
- Semantic vectors are just the matrix product of the VSM vectors and a random matrix
- The algorithm is simply reducing the dimensionality of the VSM vectors
- For certain distributions of R, R is a nearly orthogonal
- Then we can interpret this product as the projection of the VSM vectors onto a random lower dimensional space

The Random Projections Approach

Theoretical support

The Johnson and Lindenstrauss Lemma

Given $\epsilon > 0$ and an integer n, let k be a positive integer such that $k \geq k_0 = \mathrm{O}\left(\epsilon^{-2} \log n\right)$. For every set $\mathbb{P}$ of n points in $\mathbb{R}^d$ there exists $f : \mathbb{R}^d \to \mathbb{R}^k$ such that for all $u, v \in \mathbb{P}$

$$(1 - \epsilon)\|u - v\|^2 \leq \|f(u) - f(v)\|^2 \leq (1 + \epsilon)\|u - v\|^2$$

The Random Projections Approach

Theoretical support

Achlioptas Theorem

Let $\mathbb{P}$ be an arbitrary set of n points in R^d represented as an $n \times d$ matrix A. Given $\epsilon, \beta > 0$ let $k_0 = \frac{4+2\beta}{\frac{\epsilon^2}{2} - \frac{\epsilon^3}{3}} \log n$. For integer $k \geq_0$, let $\mathbb{R}$ be a $d \times k$ random matrix with $R(i,j) = r_{ij}$, where $\{r_{ij}\}$ are independent random variables from either one of the following two probability distributions:

$$r_{i,j} = \begin{cases} +1 & \text{with probability } \frac{1}{2} \\ -1 & \text{with probability } \frac{1}{2} \end{cases} \qquad r_{i,j} = \sqrt{3} \times \begin{cases} +1 & \text{with probability } \frac{1}{6} \\ 0 & \text{with probability } \frac{2}{3} \\ -1 & \text{with probability } \frac{1}{6} \end{cases}$$

Let $E = \frac{1}{\sqrt{k}} AR$. Let $f : \mathbb{R}^d \to \mathbb{R}^k$ map the i^{th} row of A to the i^{th} row of E. With probability at least $1 - n^{-\beta}$, $forallu, v \in \mathbb{P}$

$$(1-\epsilon)\|u-v\|^2 \leq \|f(u) - f(v)\|^2 \leq (1+\epsilon)\|u-v\|^2$$

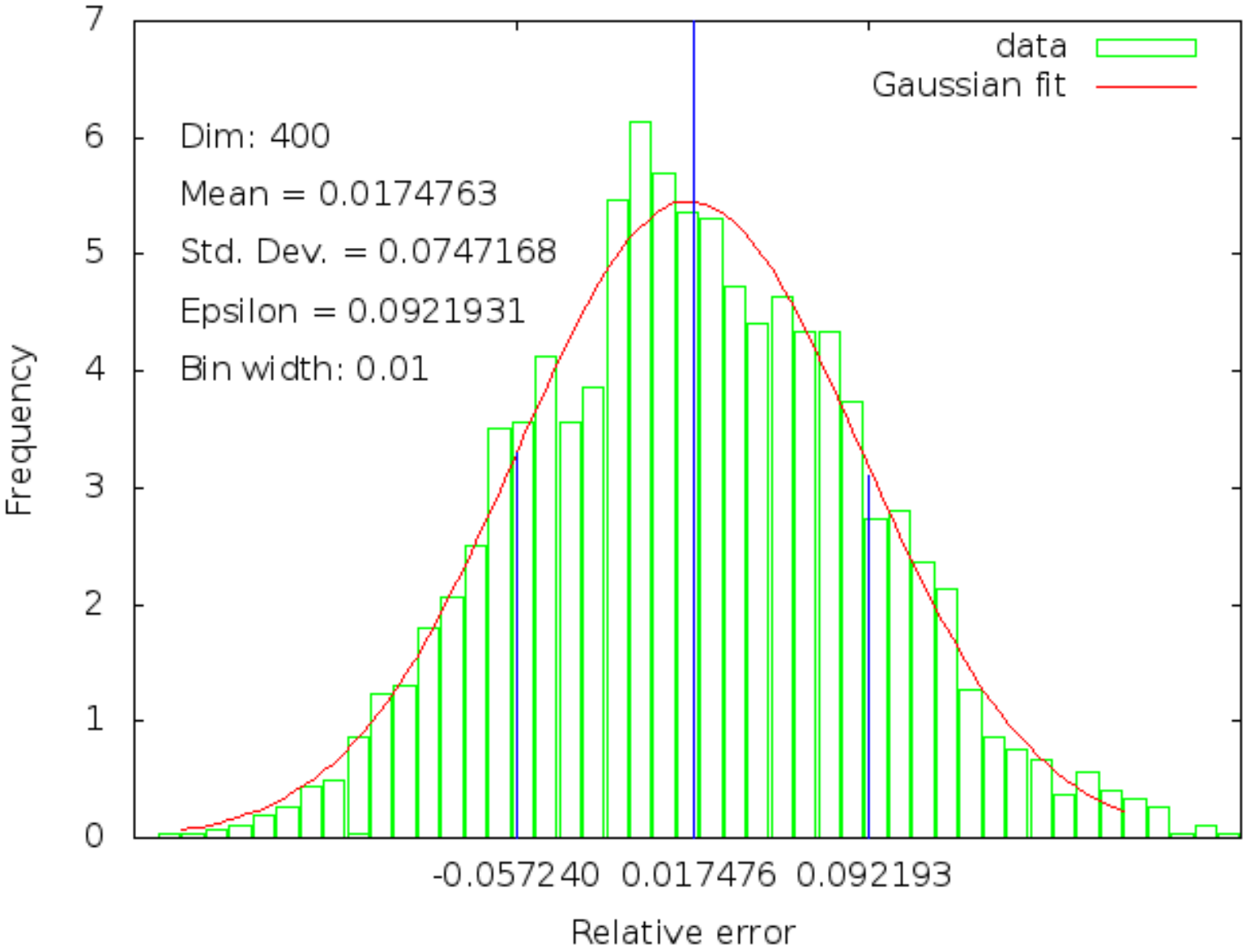
Empirical Study

Set up

- Natural language corpus with 140000 unique terms
- Filtered out low frequency terms, $n = 42905$
- Plain VSM vectors (original space is represented by the co-occurrence matrix)
- Evaluated random space dimension (k_0) ranging from 100 to 2000
- Random space generated with the sparse Achlioptas distribution
- Pick two random vectors u, v, compute squared euclidean distance $\|u - v\|^2$ and $\|f(u) - f(v)\|^2$
- Compute relative error as $\frac{\|f(u)-f(v)\|^2-\|u-v\|^2}{\|u-v\|^2}$
- Sample size: 3000

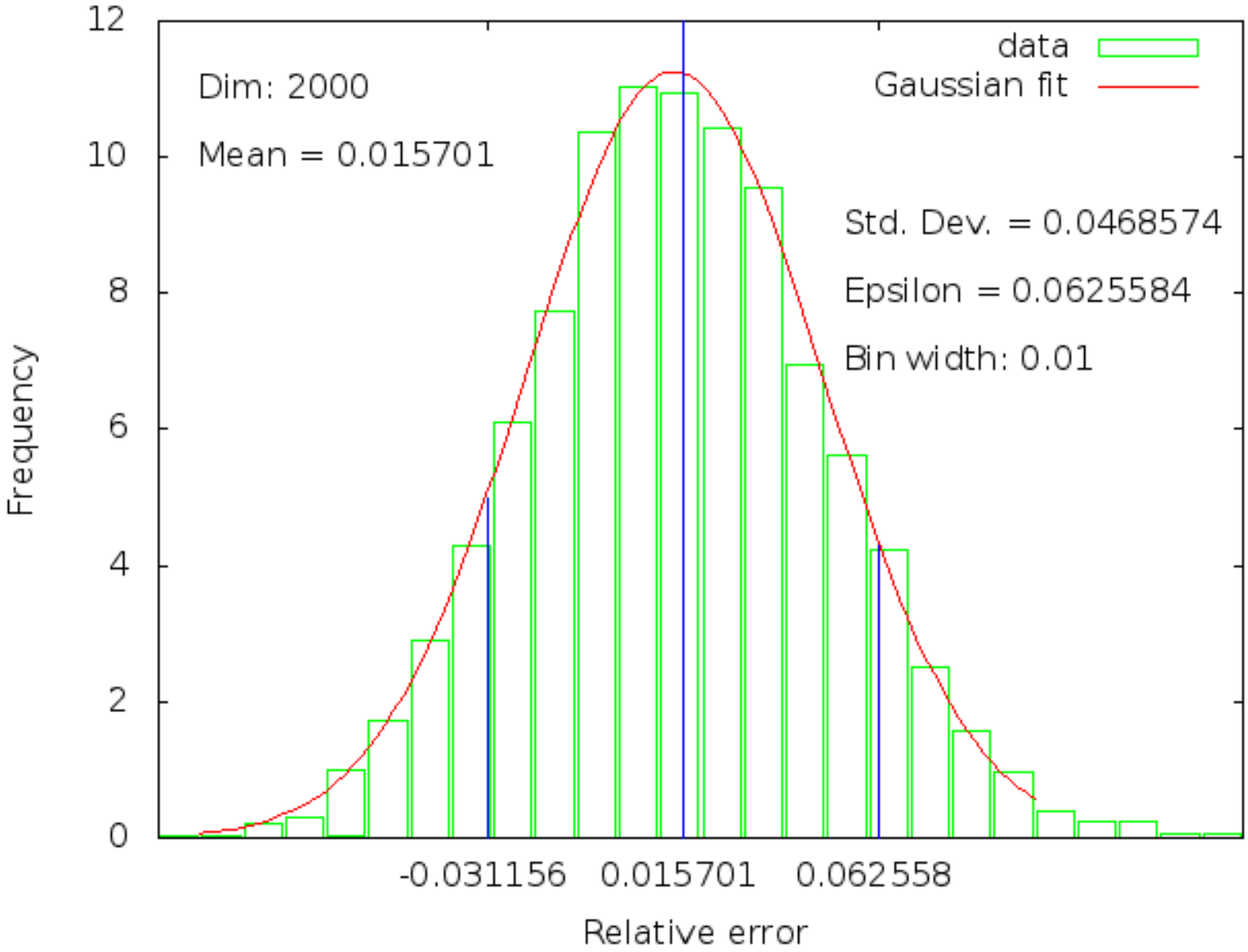
Empirical Study

Results



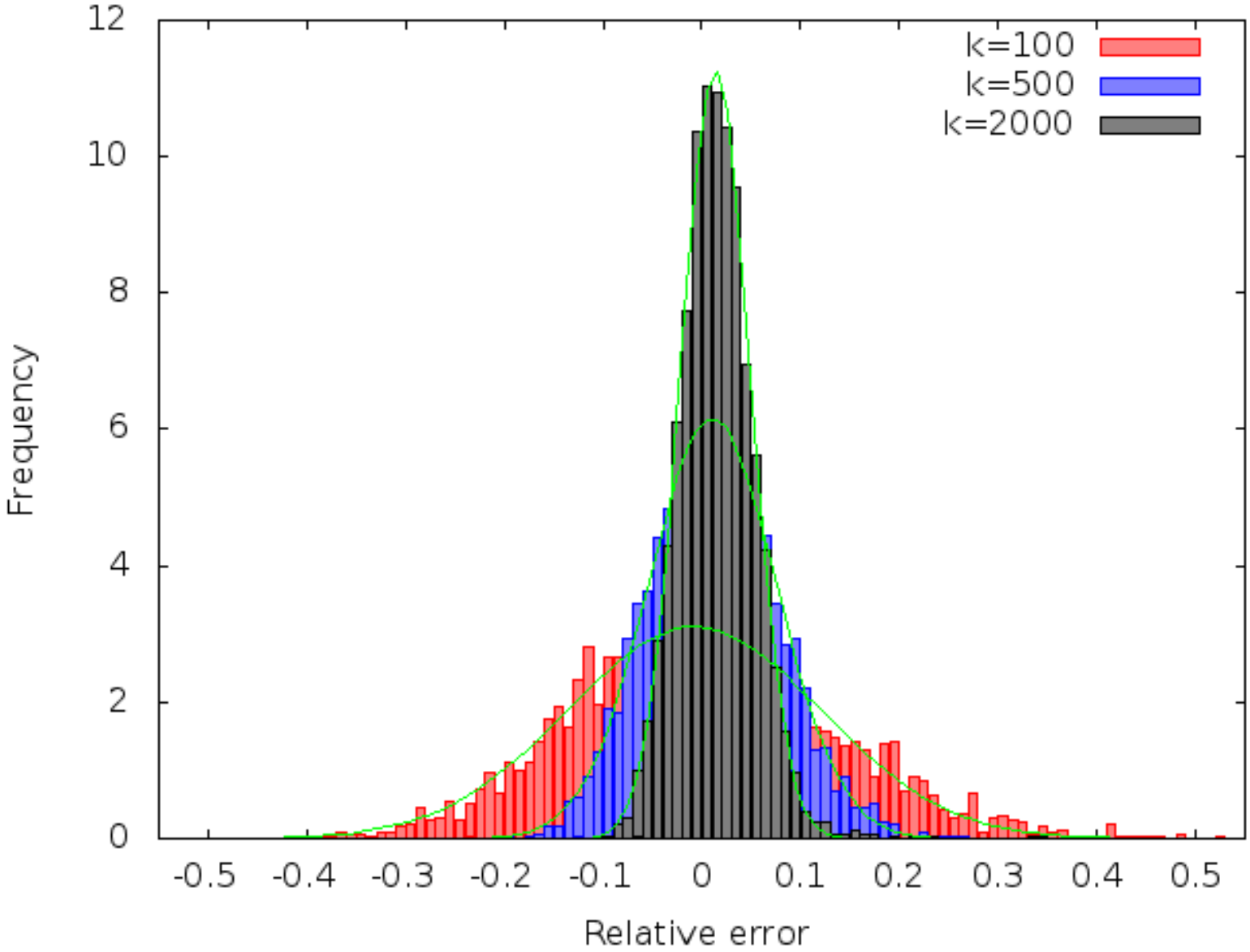
Empirical Study

Results



Empirical Study

Results



Empirical Study

Measuring ϵ

How can we compute ϵ from these histograms?

$$(1-\epsilon)\|u-v\|^2 \leq \|f(u)-f(v)\|^2 \leq (1+\epsilon)\|u-v\|^2$$

$$1-\epsilon \leq \frac{\|f(u)-f(v)\|^2}{\|u-v\|^2} \leq 1+\epsilon$$

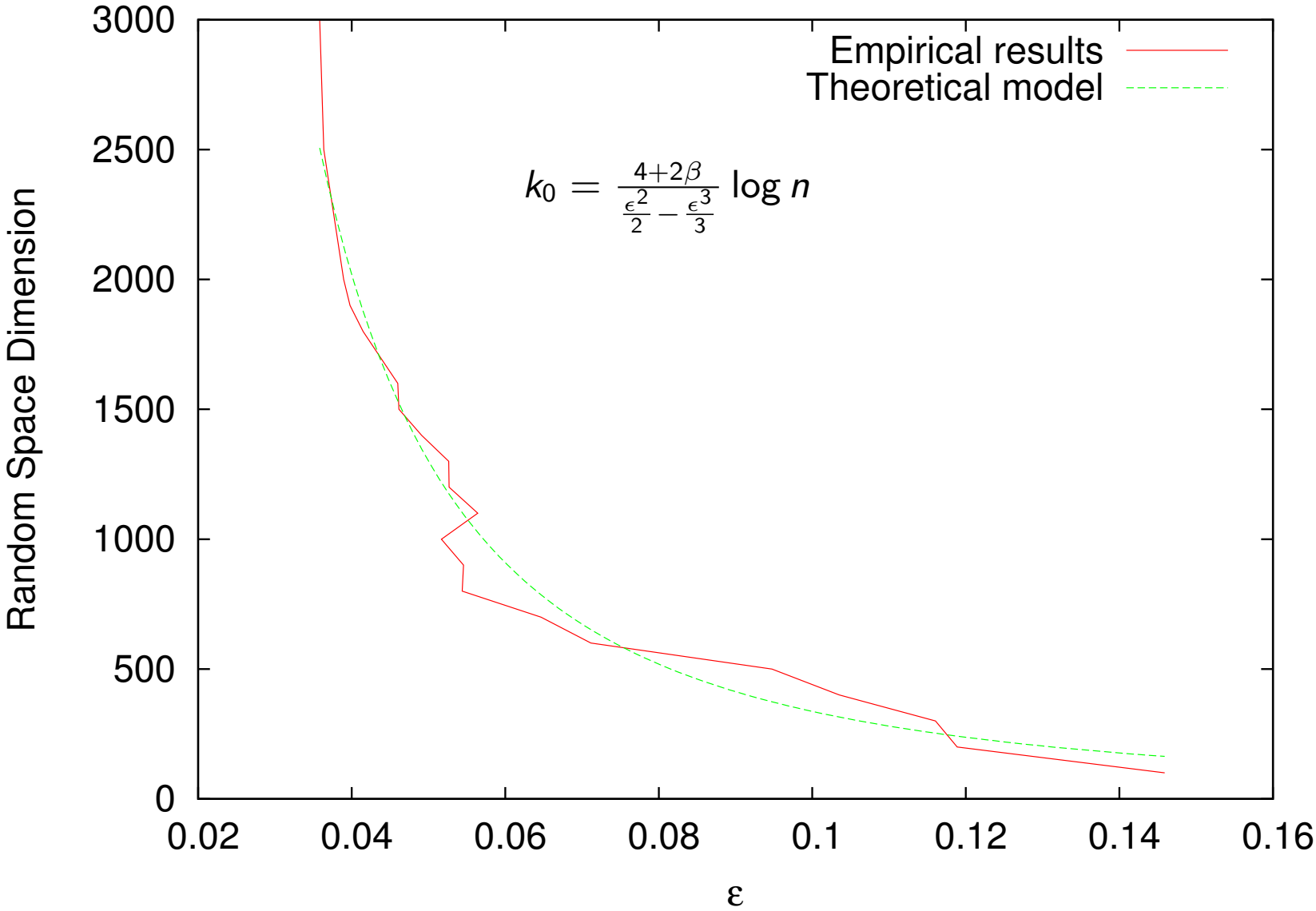
$$-\epsilon \leq \frac{\|f(u)-f(v)\|^2}{\|u-v\|^2} - 1 \leq \epsilon$$

$$-\epsilon \leq \frac{\|f(u)-f(v)\|^2-\|u-v\|^2}{\|u-v\|^2} \leq \epsilon$$

$$\epsilon = \text{máx}(|\mu-\sigma|, |\mu+\sigma|)$$

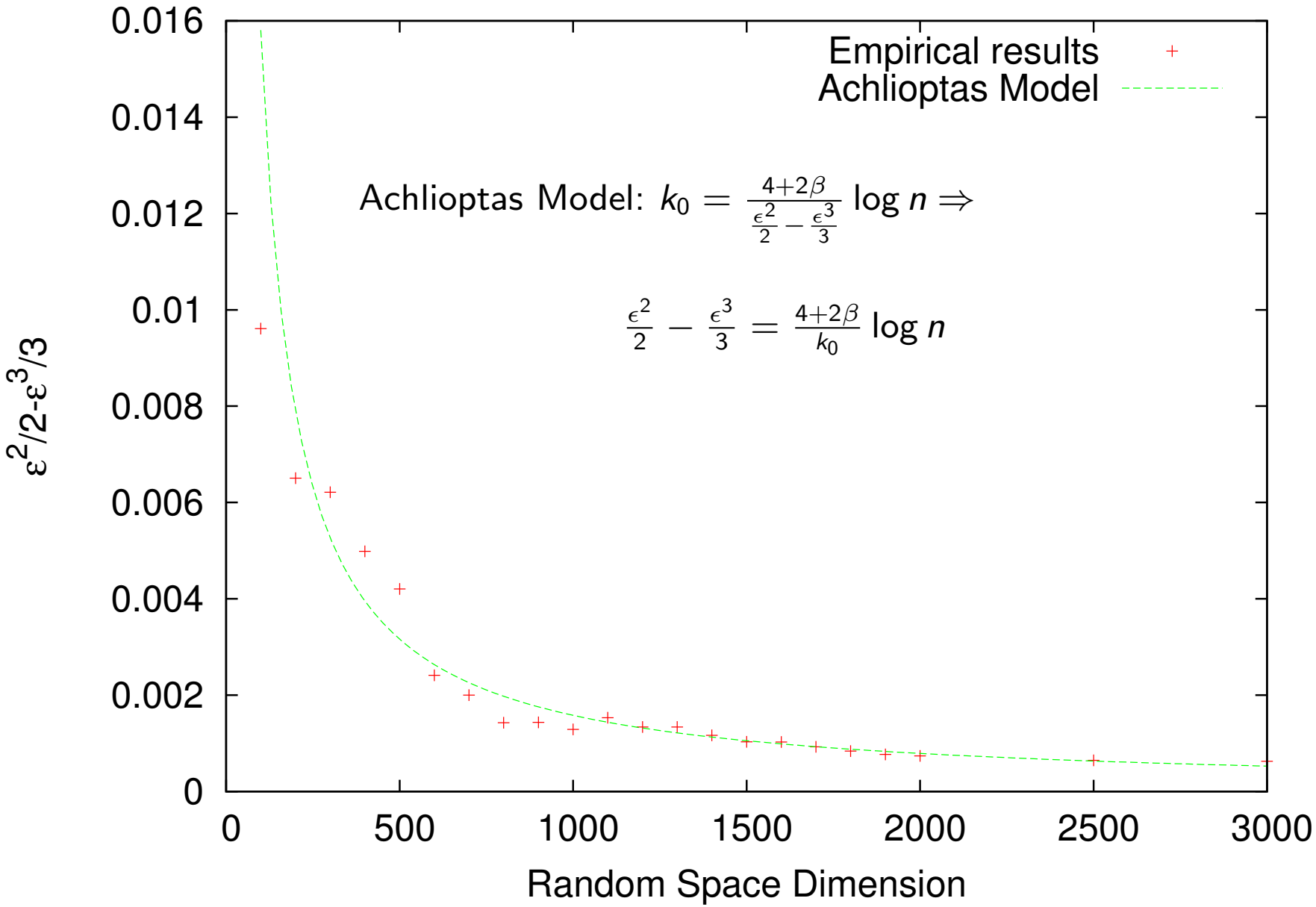
Empirical Study

Results



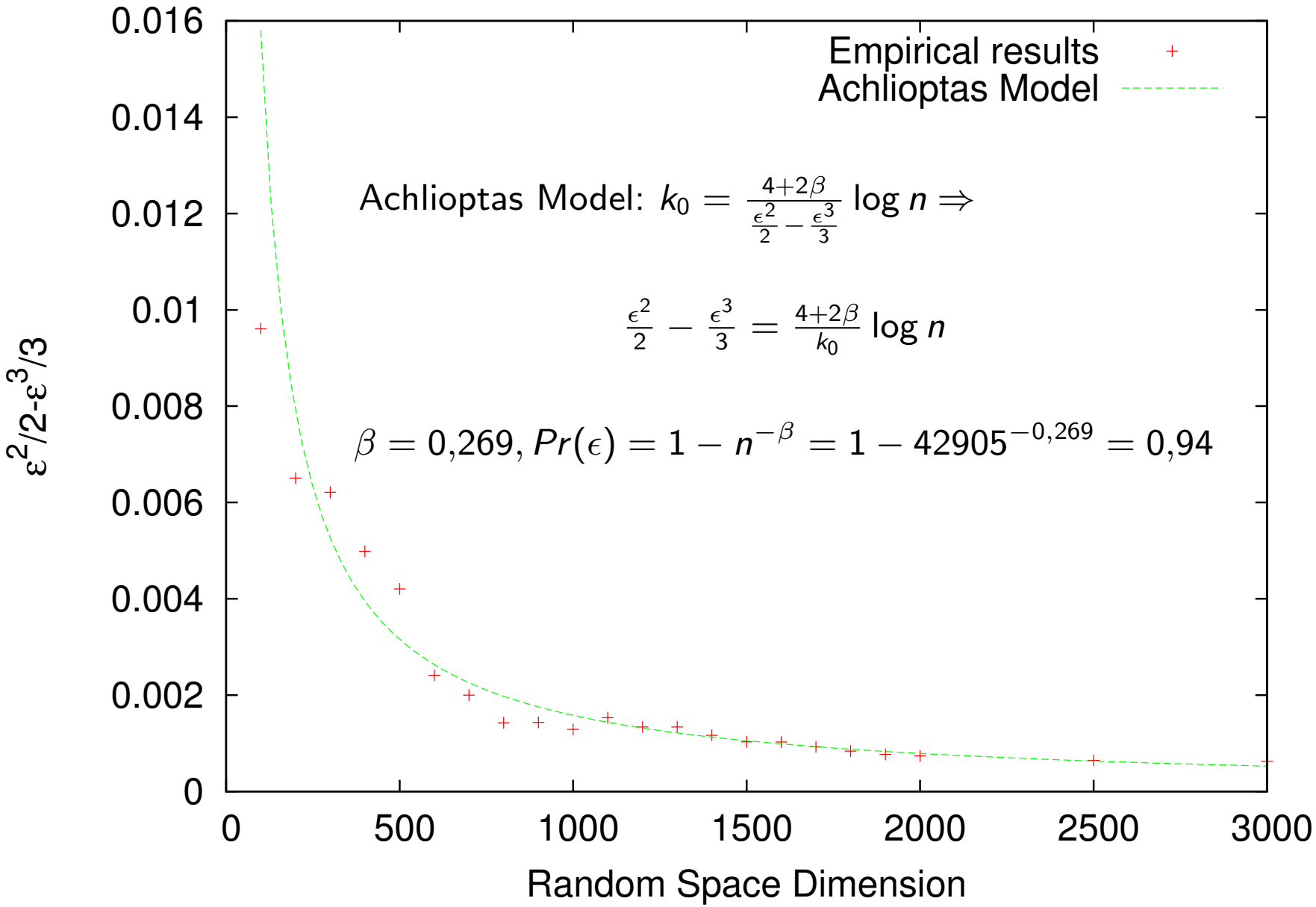
Empirical Study

Computing β



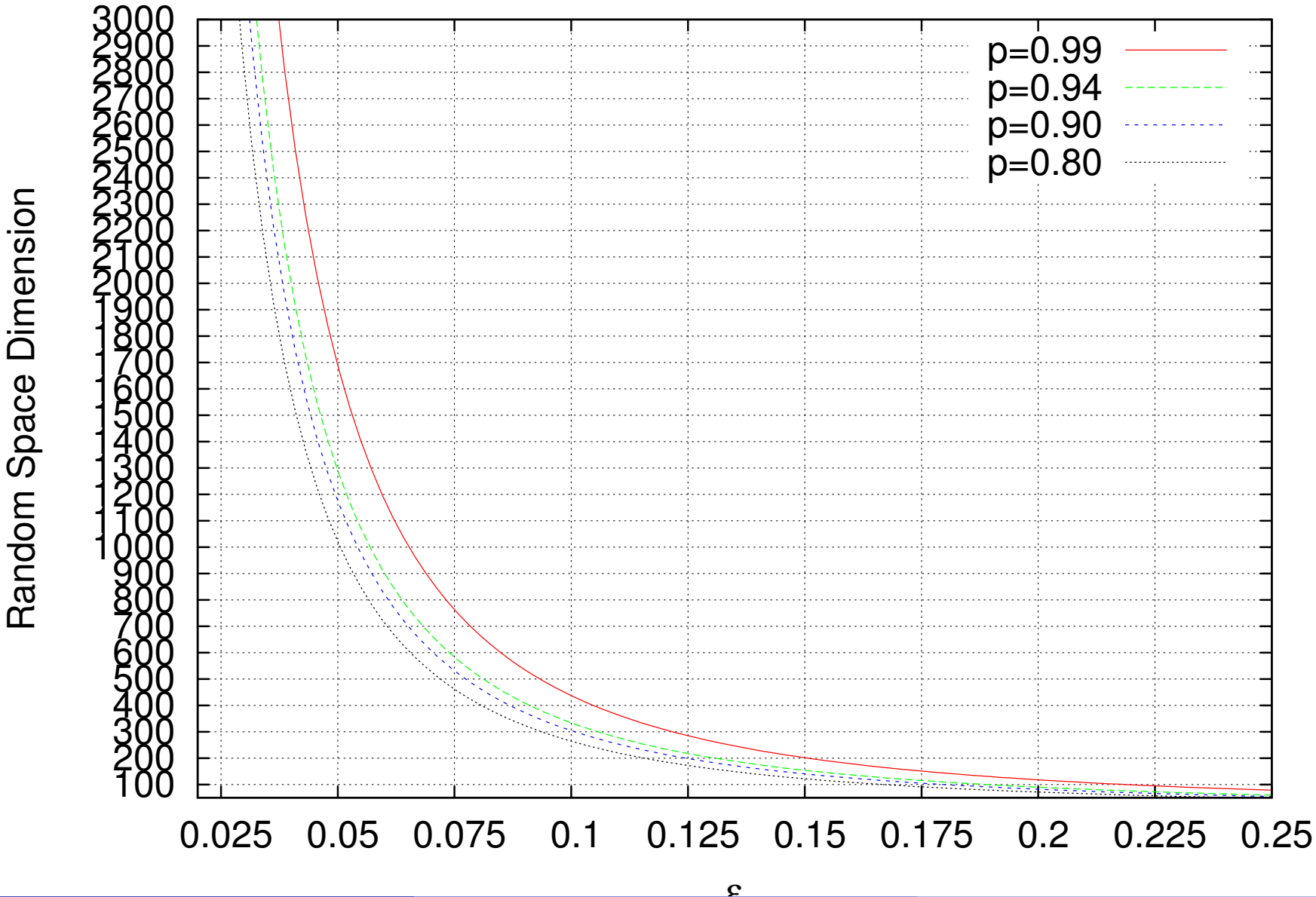
Empirical Study

Computing β



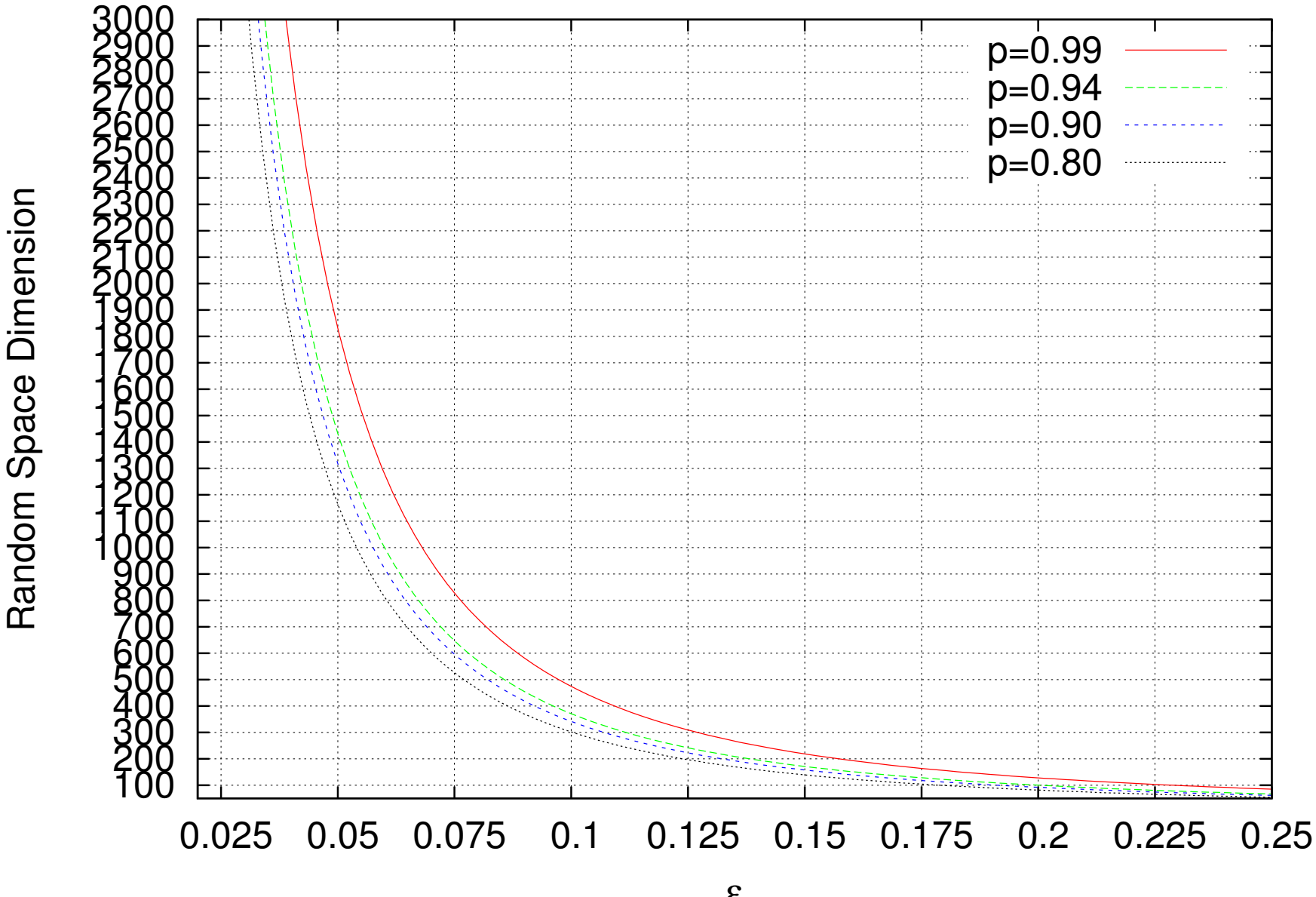
Empirical Study

Trade offs



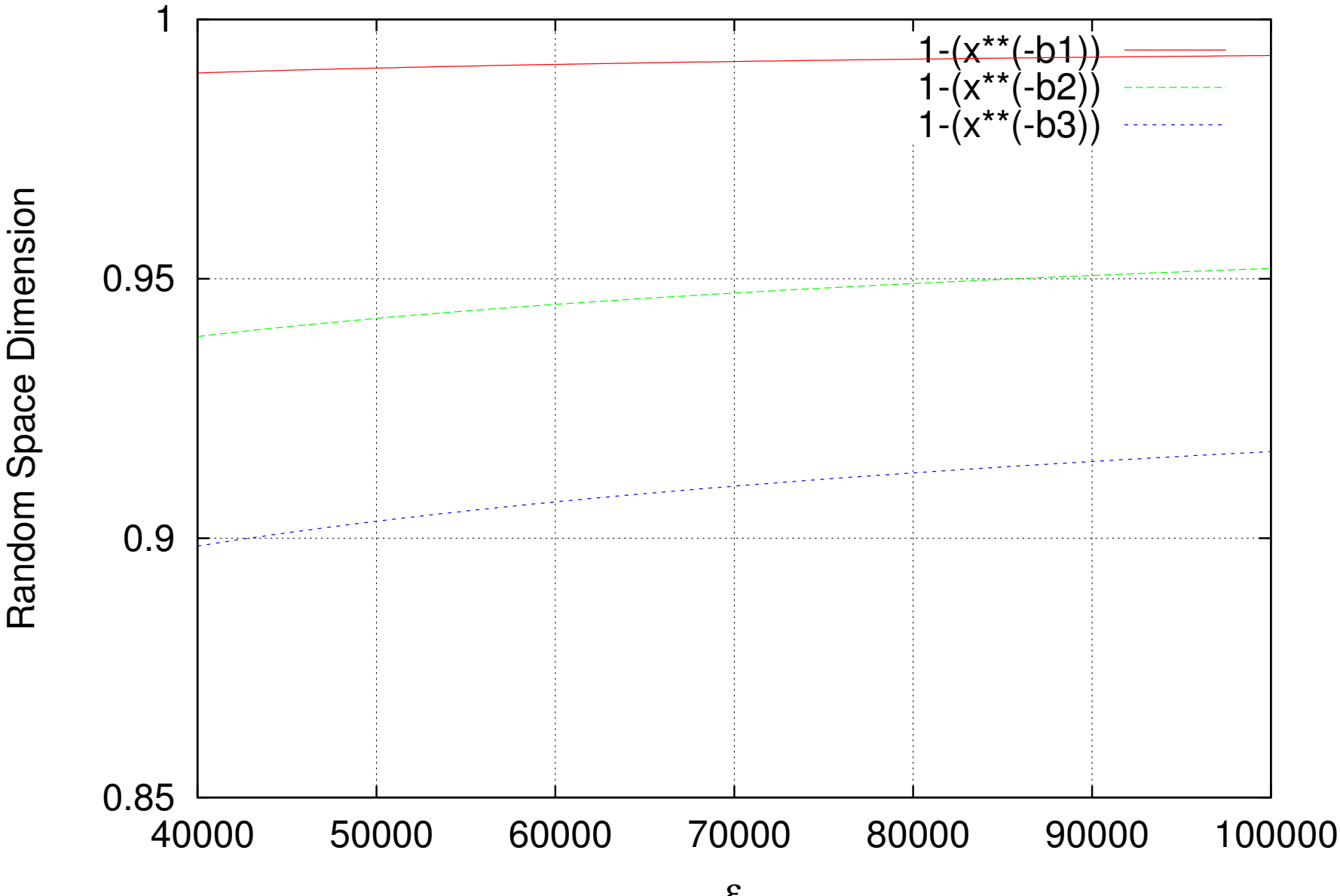
Empirical Study

Trade offs



Empirical Study

Trade offs



Empirical Study

Design

- Determine how much data is available: n
- Compute β for several certainty levels (99, 94, 90 typically)
- Compute and plot $k_0(\epsilon)$ for each p level
- Choose k according to resources constraints and accepted error level
- More data? Start again, or assume β didn't change, and increase k

Resultados

Comparación

- Matrices de confusión

	Actual Match	Actual Unmatch
Det says Matched	6212 (tp)	0 (fp)
Det says Unmatched	40862 (fn)	4775173 (tn)
Prob says Matched	39628 (tp)	197 (fp)
Prob says Unmatched	7446 (fn)	4774976 (tn)

- Metricas

	Precision	Recall	F-Score
Det	1	0,22	0,36
Prob	0,99	0,84	0,91

Resultados

Conclusión

Conclusión

- A costa de menos del 1 % de precision, Prob, detecta 4 veces mas Matches correctos que Det.
- Prob puede ser mejorado en varios aspectos incrementando tanto la precision, como la capacidad de detectar matches correctos, mientras que no es claro de que forma se puede mejorar el recall de Det.

Similitud

Un enfoque intuitivo

Idea general

Fichas similares, representan al mismo hotel, fichas disímiles representan hoteles diferentes.

- Definiciones:
  - $\mathbb{F}$: conjunto de todas las fichas de hoteles
  - $\mathbb{D}$: $\{Match, Unmatch, Review\}$ (conjunto de decisiones posibles)
  - $L : (a \in \mathbb{F}, b \in \mathbb{F}) \rightarrow \mathbb{D}$ (función decisión)
- Se construye una función de similitud entre fichas de hoteles: $S : (a \in \mathbb{F}, b \in \mathbb{F}) \rightarrow \mathbb{R}$
- S es una función escalar, cuando mayor es el valor de la función, mas similares son sus argumentos.
- Se determinan umbrales $m, u \in \mathbb{R}, m \geq u$ de similitud:

$$L(a,b) = \begin{cases} Match & \text{si } S(a,b) > m \\ Review & \text{si } u \leq S(a,b) \leq m \\ Unmatch & \text{si } S(a,b) < u \end{cases}$$

Similitud

Limitaciones

- La similitud entre fichas no necesariamente es un buen indicador de matching: dos fichas pueden ser muy similares y representar hoteles diferentes.
- Imperfecciones en la función de similitud: para dos fichas a y b muy similares $S(a, b)$ es pequeño.
- La función de similitud es naturalmente vectorial. Convertirla en escalar inevitablemente introduce ruido.

Fellegi y Sunter

Un modelo probabilístico

Idea general

Se estudia la distribución de la similitud en Matches y Unmatches.

- Valores de similitud con frecuencia alta en Matches, pero baja en Unmatches: Match.
- Valores de similitud con frecuencia alta en Unmatches, pero baja en Matches: Unmatch.
- Valores de similitud con la misma frecuencia en Matches y Unmatches: Review.

Definiciones:

- $\mathbb{P} : \mathbb{F} \times \mathbb{F}$ producto cartesiando de $\mathbb{F}$
- $\mathbb{M} : \{(a, b) \in P \mid a \text{ representa el mismo hotel que } b\}$
- $\mathbb{U} : \{(a, b) \in P \mid a \text{ no representa el mismo hotel que } b\}$
- $\mathbb{P} = \mathbb{U} \cup \mathbb{M}, \mathbb{U} \cap \mathbb{M} = \emptyset$

Fellegi y Sunter

Un modelo probabilístico: Similitud

- La función similitud es vectorial: $\gamma : (a \in \mathbb{F}, b \in \mathbb{F}) \rightarrow \mathbb{X}^n$
- Cada componente de la función γ es una función de comparación sobre un campo de la ficha (nombre del hotel, dirección, etc.)
- Estas componentes no son necesariamente escalares:
  $\gamma(a,b) =< \gamma_1(a,b), \gamma_2(a,b), ..., \gamma_n(a,b) >$
- Para cada valor posible de la función γ, se definen los siguientes parámetros:
  $m(\gamma) = P(\gamma(a,b) = \gamma|$ a representa el mismo hotel que b $)$
  $u(\gamma) = P(\gamma(a,b) = \gamma|$ a no representa el mismo hotel que b $)$
- $m(\gamma_0)$ nos dice que tanto podemos sospechar de un Match al observar γ_0
- $u(\gamma_0)$ nos dice que tanto podemos sospechar de un Unmatch al observar γ_0

Fellegi y Sunter

Un modelo probabilístico: Regla de decisión

- Los parámetros m y u se combinan en un único peso: $w(\gamma_0) = m(\gamma_0)/u(\gamma_0)$
- A mayor w mayor certeza de estar frente a un Match.
- Se determinan umbrales de peso w_m, w_u, para tomar las decisiones:

$$L(a,b) = \begin{cases} Match & \text{si } w(\gamma(a,b)) > w_m \\ Review & \text{si } w_u \leq w(\gamma(a,b)) \leq w_m \\ Unmatch & \text{si } w(\gamma(a,b)) < w_u \end{cases}$$

- Esta regla de decisión minimiza la región de revisión humana.
- La estimación de parámetros m y u se realizó mediante Maximum Likelihood Estimation.
- Los umbrales fueron determinados por cross validation.

Fellegi y Sunter

Mejoras

- La regla de decisión es muy robusta a imperfecciones en la función de similitud
- A mejor función de similitud mejor rendimiento
- Los pesos $w = m/u$ pueden calcularse de otra manera (minimizar el error)
- Los parámetros u y m pueden estimarse son métodos no supervisados
- La regla de decisión puede ser muy diferente (clasificación lineal general)

Fellegi y Sunter

Ilustración

Evaluation set

Datos históricos: pares matcheados, pares unmatcheados

- Problemas
  - Volumen insuficiente
  - Sesgados: solo pares con determinadas (pero desconocidas) características de similitud
  - Mucho ruido (falsos matches, falsos unmatches)
  - Proporción de matches y unmatches no realista
- Soluciones
  - Reduccion de ruido: ejecuciones iniciales del modelo probabilístico determinaron candidatos a revisión humana que fueron corregidos
  - Incremento de volumen, reducción de sesgo: clausura bajo transitividad de matches
  - Incremento de volumen, reducción de sesgo, corrección de proporción entre clases: complemento a producto cartesiano.
  - El complemento a producto cartesiano produce ruido sistemático por generar falsos unmathces
  - Reduccion de ruido sistemático: Fellegi Sunter sobre clases de equivalencia.

Referencias

Ivan P. Fellegi y Alan B. Sunter

A Theory for Record Linkage.

Journal of the American Statistical Association Volume 64, Issue 328, 1969.

S. Deerwester, et al.

Indexing by latent semantic analysis.

JASIS 41.6 (1990): 391-407.

P. Kanerva, et al.

Random indexing of text samples for latent semantic analysis.

Proceedings of the 22nd annual conference of the cognitive science society. Vol. 1036. 2000.

D. Achlioptas.

Database-friendly random projections: Johnson-Lindenstrauss with binary coins.

Journal of computer and System Sciences 66.4 (2003): 671-687.