Faster and Accurate Similarity Evaluations via Sampling

Conrado Martínez
Univ. Politècnica de Catalunya, Barcelona, Spain

Joint work with:

Jun Wang (UPC)

Introduction

- There are many applications that involve making a huge number of distance/similarity evaluations between complex objects
- For example:
  - Locate the best match for a song or image
  - Find the best match for some DNA sample in a crime scene
  - Cluster data into groups
  - Facial recognition
  - Reconstruct a phylogenetic tree
  - ...

Introduction

Several different (and often complementary) approaches to reduce the computational cost

1. Reduce the number of distance evaluations organizing the data (vp-trees, Burkhard-Keller trees, GHTs, GNATs, ...) exploiting triangle inequality
2. Filter with a simpler distance function δ': $\delta'(A, B) > r' \implies \delta(A, B) > r$
3. Reduce the "dimensionality" using a simpler distance function with approximation guarantees

Introduction

We focus here in distance/similarity measures between sets and multisets

- In many applications we can modellize our complex object as a set or multiset
- For example,
  - from a long DNA sequence we can extract the set/multiset of k-mers
  - from a textual document we can extract the *vocabulary*, the set of distinct words (or stems) in the document
  - ...

Similarity measures

Consider two sets A and B.

Jaccard's index	$\frac{\lvert A \cap B \rvert}{\lvert A \cup B \rvert}$
Otsuka-Ochiai (a.k.a. Cosine)	$\frac{\lvert A \cap B \rvert}{\sqrt{\lvert A \rvert \cdot \lvert B \rvert}}$
Sørensen-Dice	$2\frac{\lvert A \cap B \rvert}{\lvert A \rvert + \lvert B \rvert}$
Kulczynski 1	$\frac{\lvert A \cap B \rvert}{\lvert A \triangle B \rvert}$
Kulczynski 2	$\frac{1}{2}\left(\frac{\lvert A \cap B \rvert}{\lvert A \rvert} + \frac{\lvert A \cap B \rvert}{\lvert B \rvert}\right)$
Simpson	$\frac{\lvert A \cap B \rvert}{\min(\lvert A \rvert, \lvert B \rvert)}$
Braun-Blanquet	$\frac{\lvert A \cap B \rvert}{\max(\lvert A \rvert, \lvert B \rvert)}$
Correlation	$\cos^2(A, B) = \frac{\lvert A \cap B \rvert^2}{\lvert A \rvert \cdot \lvert B \rvert}$
...	...

Estimating Similarity

Let S_A and S_B two random samples of A and B, resp., and σ a similarity measure.

- Does $\sigma(S_A, S_B)$ give us an unbiased estimation of $\sigma(A, B)$?
- If not, can we find an unbiased estimator for $\sigma(A, B)$ from the information in the samples?
- How accurate are these estimations? How do they depend on the size of the samples? $\Longrightarrow$ trade-off: large samples should lead to more accurate estimates but computationally more costly!

Sampling

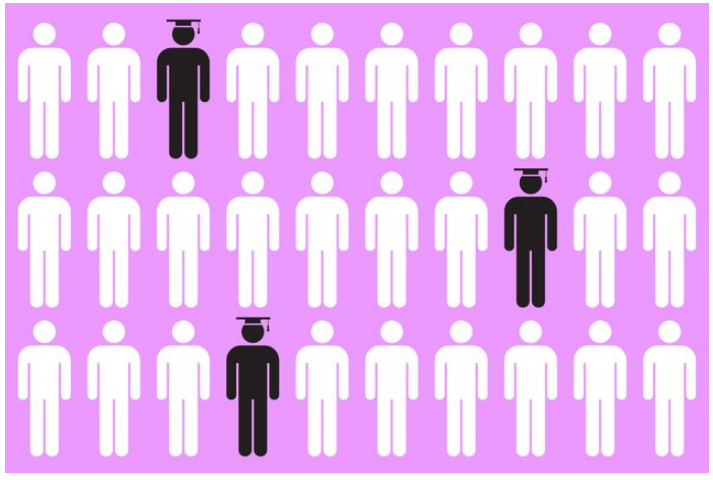
procedure DISTINCTSAMPLING(k, $\mathcal{Z}$)
    fill S with the first k distinct elements (and hash values)
    of the stream $\mathcal{Z}$
    for all $z \in \mathcal{Z}$ **do**
        if HASH(z) $<$ min hash value in S **then**
            Discard z; **continue**
        ▷ HASH(z) $>$ min hash value in S
        if $z \in S$ **then**
            Update z stats
        else ▷ replace elem of min. hash with z
            $S \leftarrow S \setminus \{\text{elem. with min. hash in } S\} \cup \{z\}$
    return S

Distinct Sampling

- The algorithm draws a random sample of k distinct elements (each one has prob. $1/n$ of being drawn, $n =$ number of distinct elements), by keeping in the sample the k elements with the largest hash values seen so far[1]
- If we use uniform random numbers in $(0, 1)$ instead of hash values (and don't check if $z \in S$) $\Rightarrow$ Reservoir Sampling

[1] Pragmatic assumptions: hash values uniformly distributed; probability of collisions negligible

Affirmative Sampling (Lumbroso & M., 2019)



- The larger the cardinality (n) the larger the samples $\Rightarrow$ samples better represent diversity
- All distinct elements have the same opportunity to be sampled

Affirmative Sampling

procedure AFFIRMATIVESAMPLING(k, $\mathcal{Z}$)
    fill S with the first k distinct elements (and hash values)
    of the stream $\mathcal{Z}$
    for all $z \in \mathcal{S}$ **do**
        if HASH(z) $<$ min hash value in S **then**
            Discard z; **continue**
        ▷ HASH(z) $>$ min hash value in S
        if $z \in S$ **then**
            Update z stats
        else if HASH(z) $>$ k-th largest hash value in S **then**
            $S \leftarrow S \cup \{z\}$
        else ▷ replace elem of min. hash with z
            $S \leftarrow S \setminus \{$elem. with min. hash in $S\} \cup \{z\}$
    return S

Affirmative Sampling

- The size of the sample S is a random variable = the number of k-records in a random permutation of size n
- The sample does not contain the k-records, but the $|S|$ elements with the largest hash values seen so far $\Rightarrow$ S is a random sample
- If $x \in S$ then x has been added to S in its very first occurrence and it has remained in S ever since $\Rightarrow$ can collect exact stats (e.g. frequency counts) for x

Affirmative Sampling

- Properties of $|S|$ are very well understood; in particular

$$\mathbb{E}\{|S|\} = k \ln(n/k) + \text{l.o.t.}$$

  The exact and asymptotic distribution of R, moments, ... is known (e.g., Helmi, M., Panholzer, 2014)
- Estimating cardinality (RECORDINALITY, Helmi, Lumbroso, M., Viola, 2012)

$$\mathbb{E}\left\{k\left(1+\frac{1}{k}\right)^{|S|-k+1}-1\right\} = n$$

Affirmative Sampling

- We also understand fairly well F = number of times an element substitutes another in the sample (not a k-record, but larger than some k-record):

$$\mathbb{E}\{F_n\} = k \ln^2(n/k) + \text{l.o.t.}$$

- Expected cost of Affirmative Sampling

$$\begin{aligned}\mathbb{E}\{C\} &= \Theta(N + (\mathbb{E}\{|S|\} + \mathbb{E}\{F\}) \log \mathbb{E}\{|S|\}) \\ &= \Theta(N + (\log^2 n) \cdot (\log \log n))\end{aligned}$$

using hashing for membership and a couple of priority queues (one of fixed size k, the other of size $|S| - k$)

Estimating Proportions

Let P some property.

- $n = $ # of distinct elements
- $n_P = $ # of distinct elements that satisfy P
- $S = $ size of the sample $\Leftarrow$ in general, a r.v., assume $2 \leqslant S \leqslant n$
- $S_P = $ # of elements in the sample that satisfy P

Theorem

1. $\mathbb{E}\left\{\frac{S_P}{S}\right\} = \frac{n_P}{n}$
2. $\mathbb{V}\left\{\frac{S_P}{S}\right\} \sim \frac{n_P}{n} \cdot \left(1 - \frac{n_P}{n}\right) \cdot \mathbb{E}\left\{\frac{1}{S}\right\}$

For affirmative sampling $\mathbb{E}\{1/S\} \sim 1/\mathbb{E}\{S\} = 1/(k \ln(n/k))$

Estimating the Jaccard similarity

- S_A and S_B random samples from A and B, resp.
- For any set X let $\tau_X = \min\{\text{hash}(x) \mid x \in X\}$ and $X^{\geqslant \tau} = \{x \in X \mid \text{hash}(x) \geqslant \tau\}$
- Let $\tau = \max\{\tau_{S_A}, \tau_{S_B}\}$. Then
  1. $S_A^{\geqslant \tau} \cup S_B^{\geqslant \tau} = (S_A \cup S_B)^{\geqslant \tau}$ is a random sample of $A \cup B$
  2. $(S_A^{\geqslant \tau} \cup S_B^{\geqslant \tau}) \cap (A \cap B) = (S_A^{\geqslant \tau} \cap S_B^{\geqslant \tau}) = (S_A \cap S_B)^{\geqslant \tau} = (S_A \cap S_B)$

Estimating the Jaccard similarity

Theorem

1. $\mathbb{E}\left\{J(S_A^{\geqslant\tau}, S_B^{\geqslant\tau})\right\} = J(A,B) = \frac{|A\cap B|}{|A\cup B|}$
2. $\mathbb{V}\left\{J(S_A^{\geqslant\tau}, S_B^{\geqslant\tau})\right\} \sim \frac{J(A,B)\cdot(1-J(A,B))}{k\ln(|A\cup B|/k)}$

Estimating the size of the intersection

Moreover we can estimate the size of the intersection with:

$$Z_1 = \frac{|S_A \cap S_B|}{|S_A|} \cdot \left(k \left(1 + \frac{1}{k} \right)^{|S_A|-k+1} - 1 \right)$$

$$Z_2 = \frac{|S_A \cap S_B|}{|S_A|} \cdot \frac{|S_A| - 1}{1 - \tau_{S_A}}$$

$$\mathbb{E}\{Z_1\} = \mathbb{E}\{Z_2\} = |A \cap B|$$

N.B. No need to “filter” the sample S_A

Estimating other similarity measures

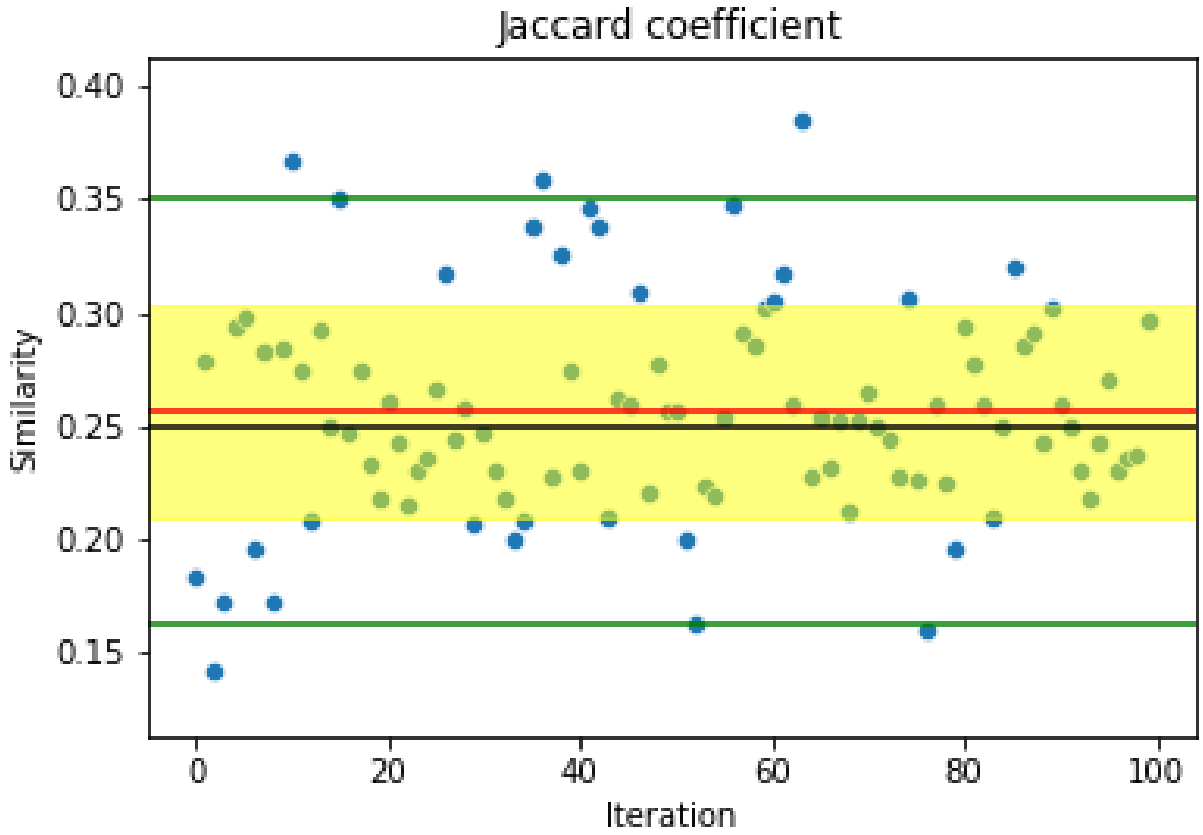
The same proof that works for Jaccard's similarity also works for:

1. Containment: $c(A, B) = |A \cap B|/|A|$ (this index is 1 if $A \subseteq B$); $\mathbb{E}\{c(S_A, S_B)\} = c(A, B)$
2. If σ is any of Jaccard, Simpson, Braun-Blanquet, Kulczynski 2 or Sørensen-Dice:

$$\mathbb{E}\left\{\sigma(S'_A, S'_B)\right\} = \sigma(A, B)$$

3. We conjecture this also holds (asymptotically) for cosine, correlation and Kulczynski 1

A few simulations



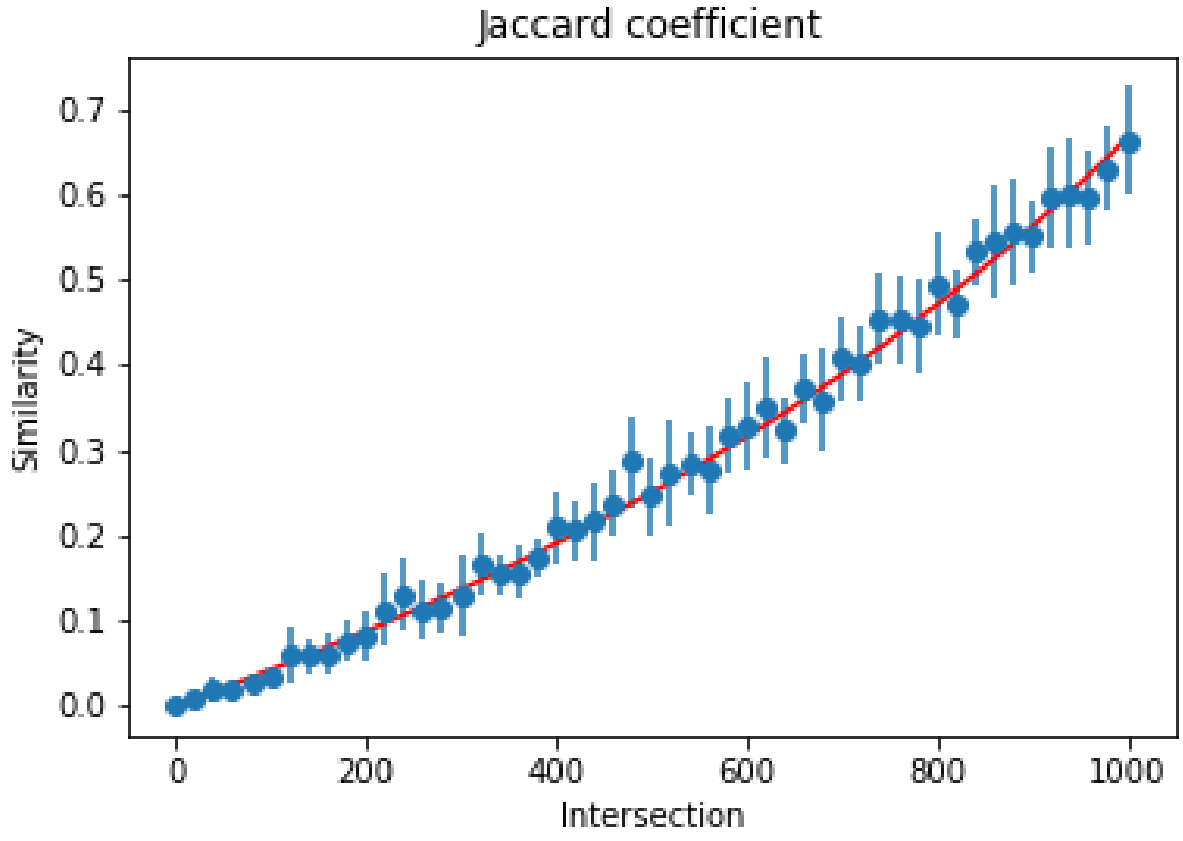
$|A| = 1000, |B| = 1500, J(A, B) = 0.25$

red line = avg. of 100 experiments

yellow band = avg $\pm$ std. deviation

green lines = 95% of observations

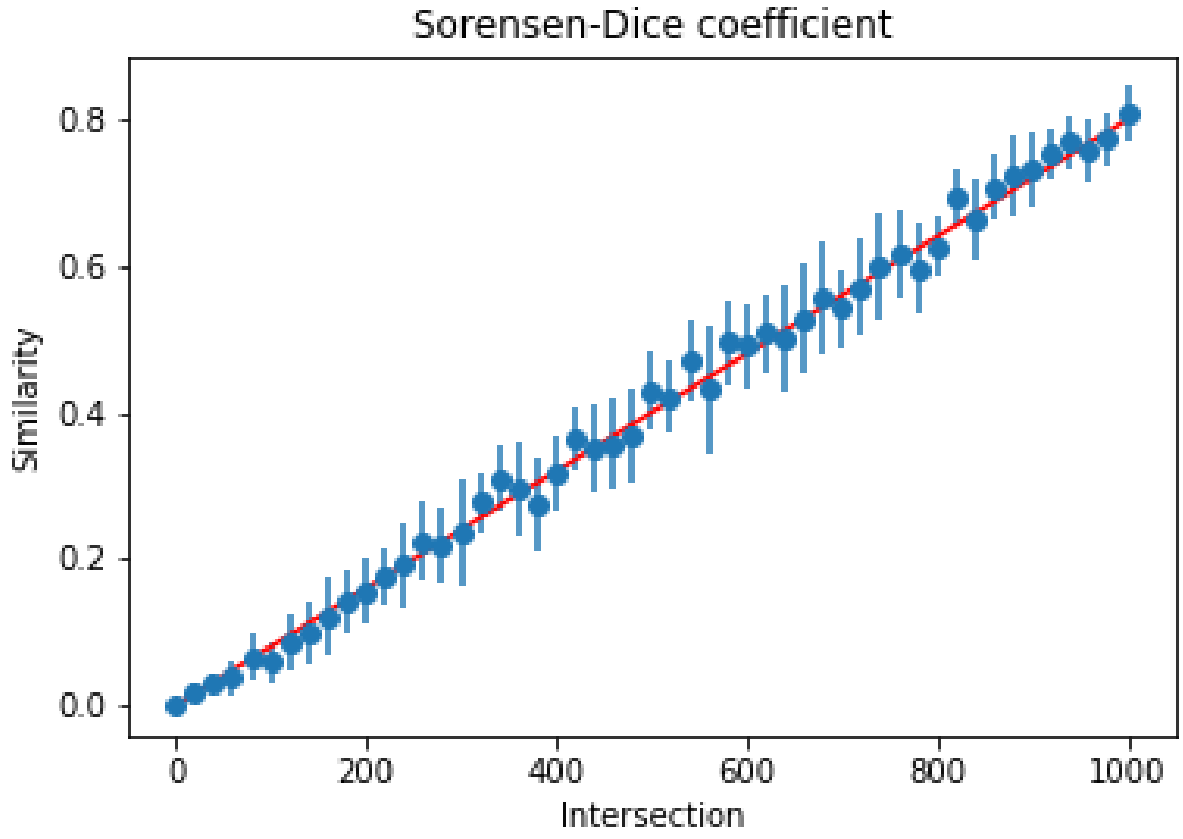
A few simulations



$|A| = 1000, |B| = 1500, |A \cap B| \in \{0, 10, \ldots, 1000\}$

red line = $J(A, B) = |A \cap B|/|A \cup B|$

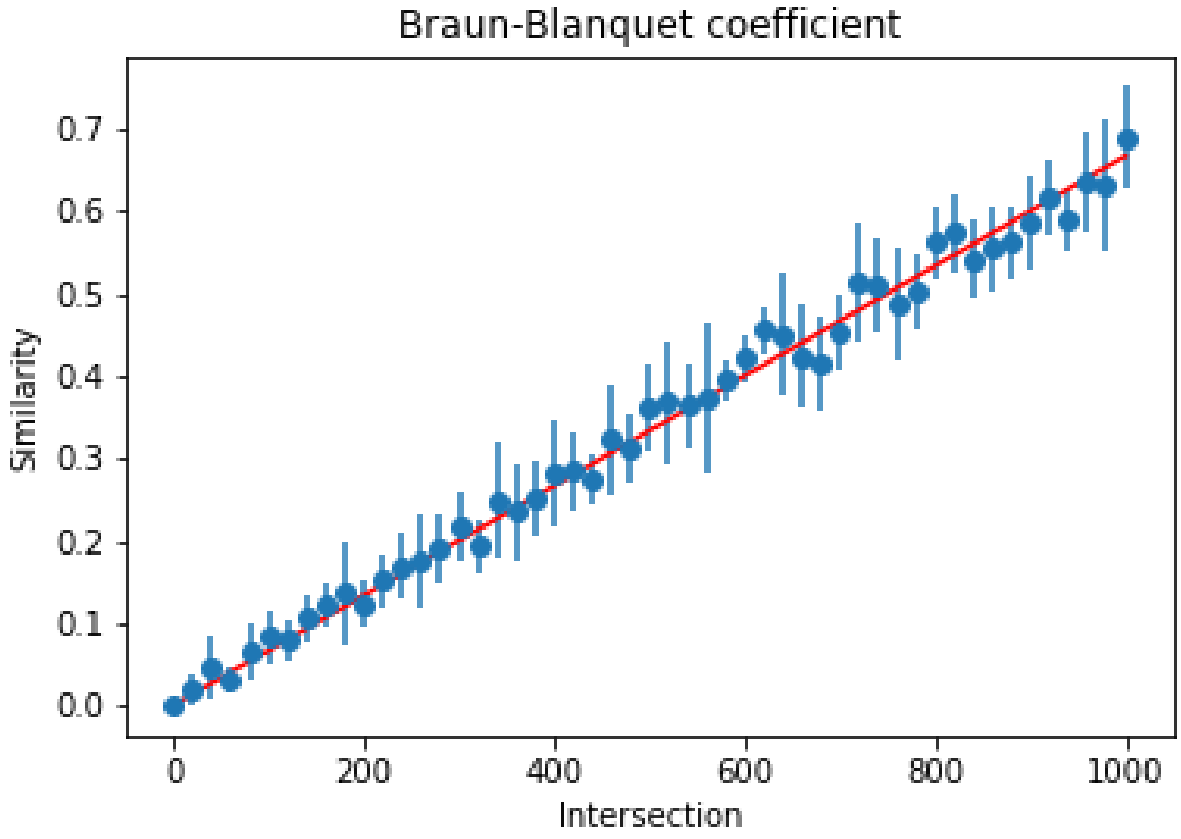
A few simulations



$|A| = 1000, |B| = 1500, |A \cap B| \in \{0, 10, \ldots, 1000\}$

red line = Sørensen-Dice$(A, B) = 2|A \cap B/(|A| + |B|)$

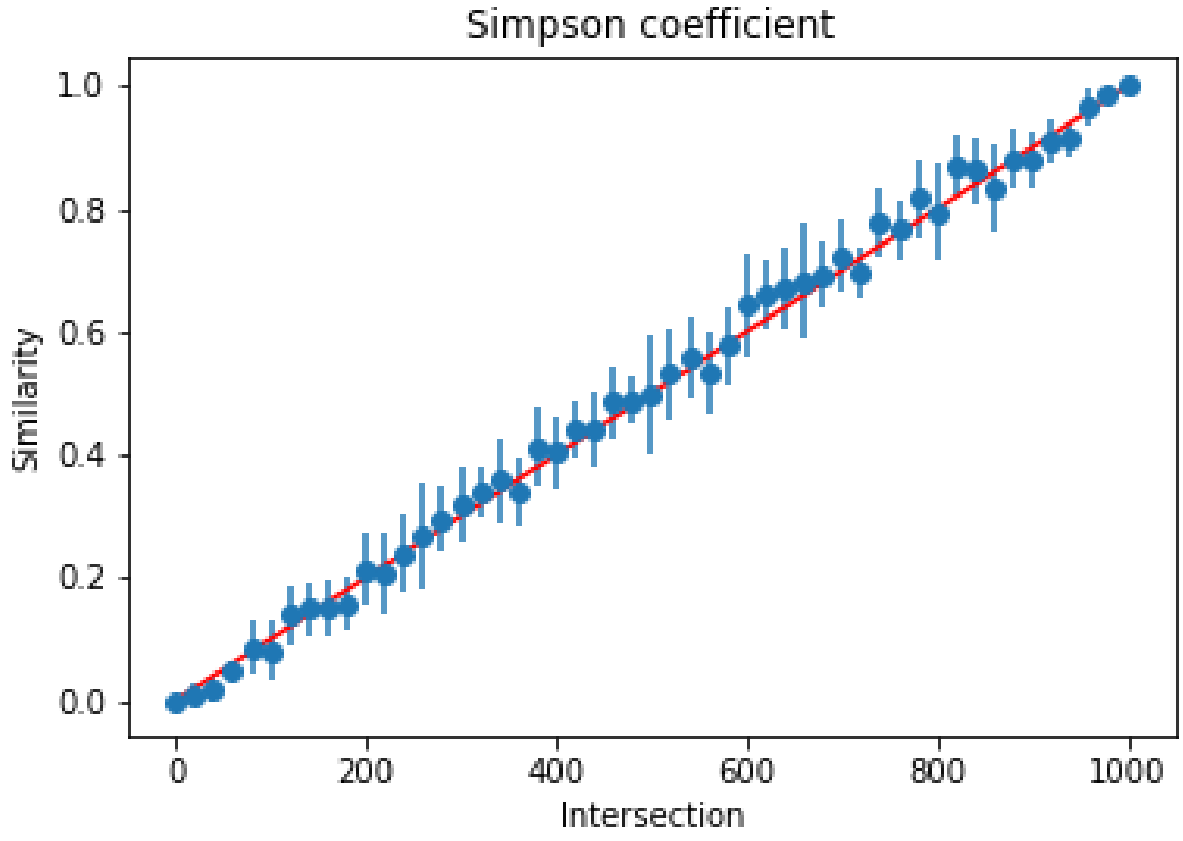
A few simulations



$|A| = 1000, |B| = 1500, |A \cap B| \in \{0, 10, \ldots, 1000\}$

red line = Braun-Blanquet$(A, B) = |A \cap B| / \max(|A|, |B|)$

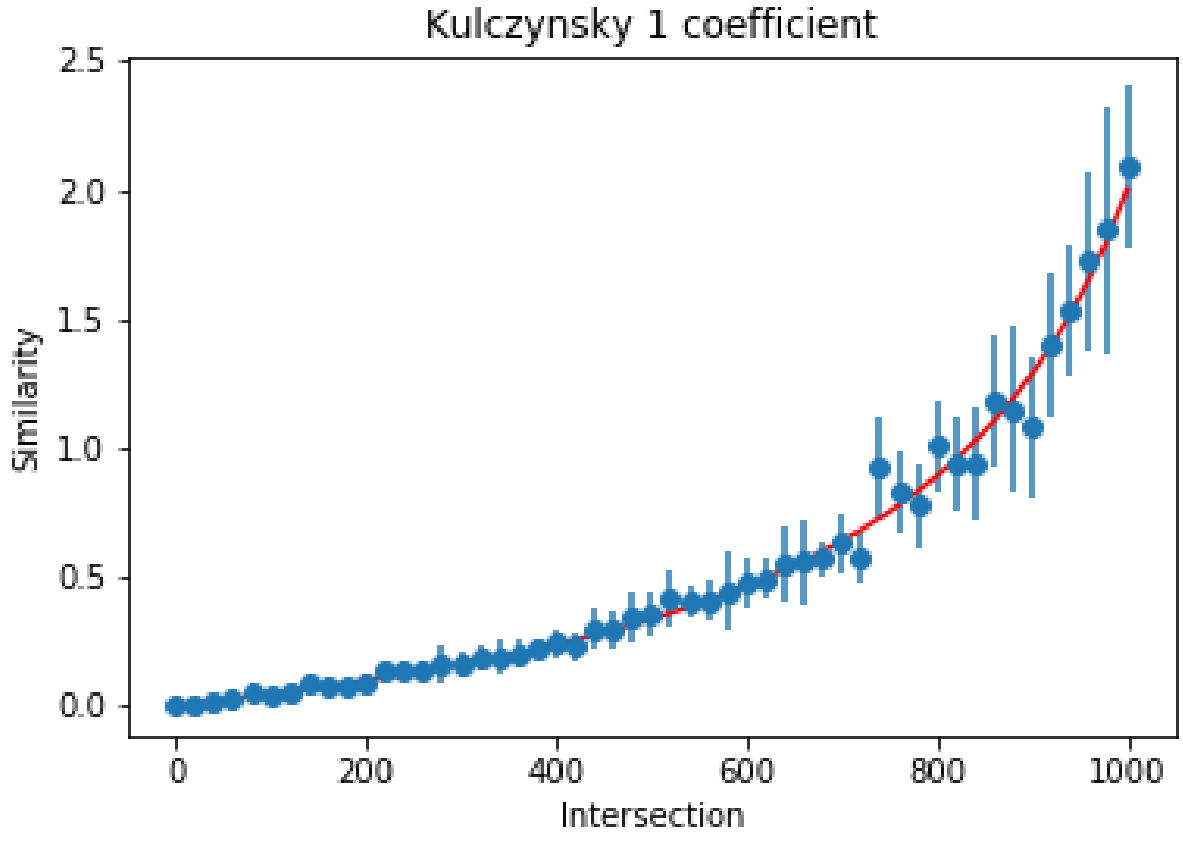
A few simulations



$|A| = 1000, |B| = 1500, |A \cap B| \in \{0, 10, \ldots, 1000\}$

red line = Simpson$(A, B) = |A \cap B| / \min((|A|, |B|)$

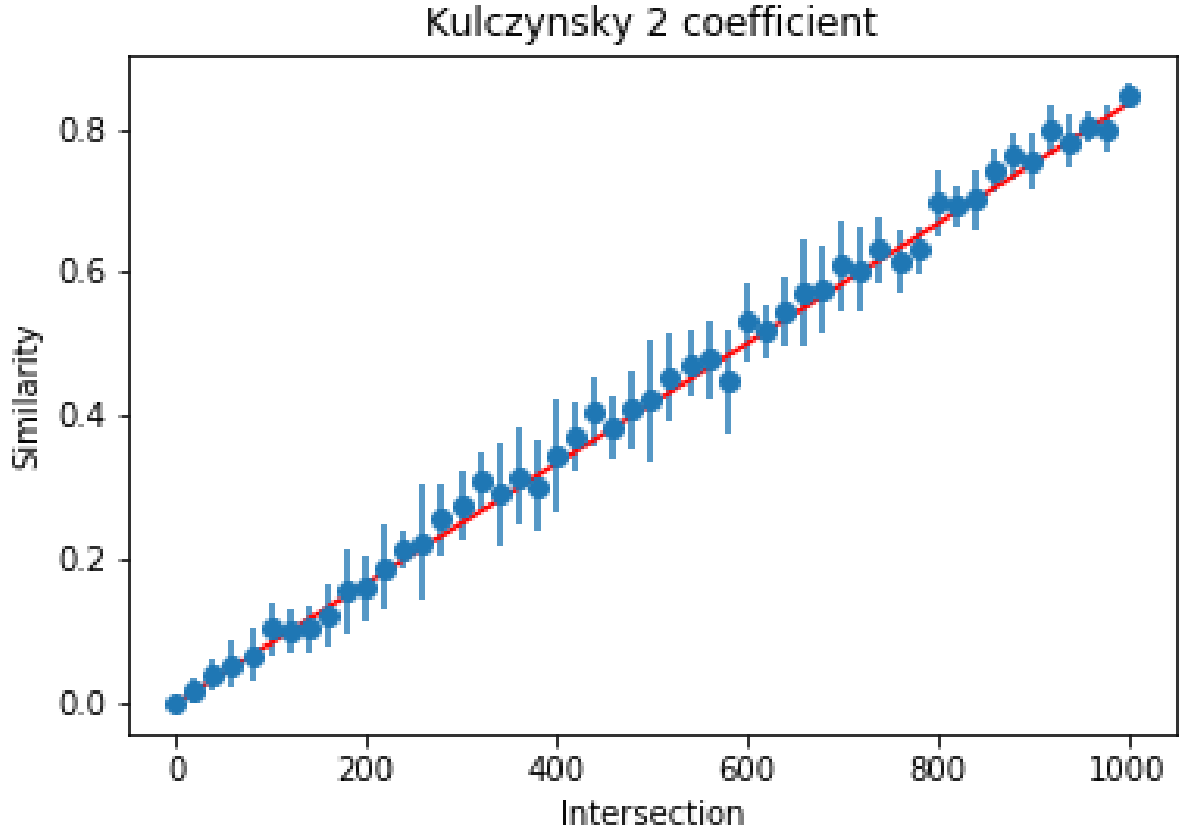
A few simulations



$|A| = 1000, |B| = 1500, |A \cap B| \in \{0, 10, \ldots, 1000\}$

red line = $K_1(A, B) = |A \cap B|/|A \triangle B|$

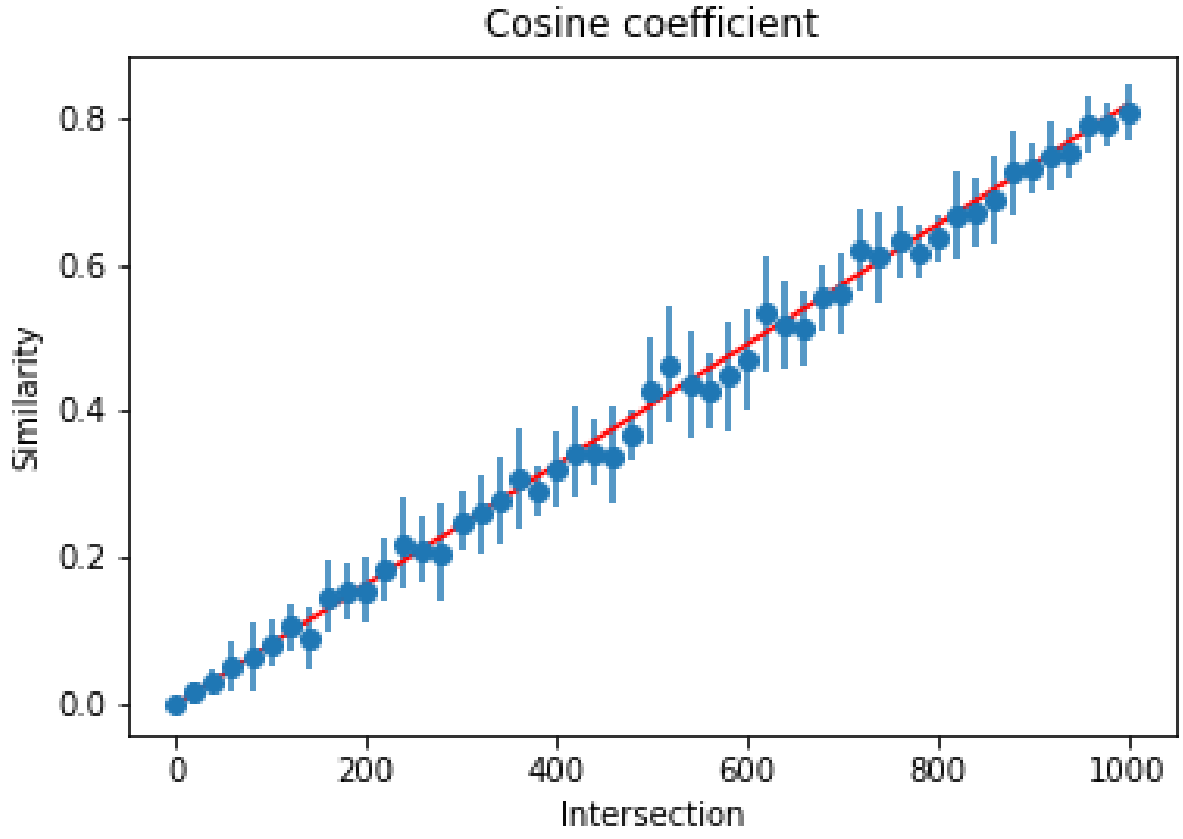
A few simulations



$|A| = 1000, |B| = 1500, |A \cap B| \in \{0, 10, \ldots, 1000\}$

red line = $K_2(A, B) = \frac{1}{2}\left(|A \cap B|/|A| + |A \cap B|/|B|\right)$

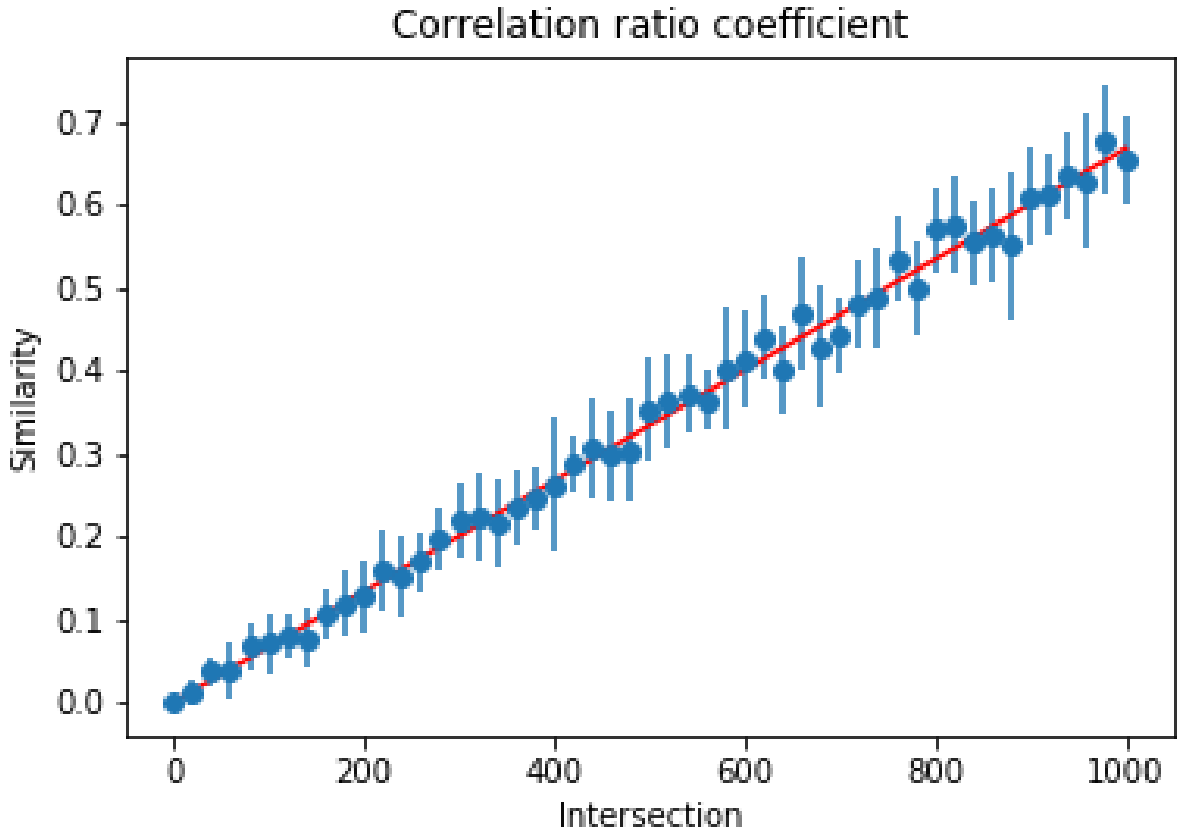
A few simulations



$|A| = 1000, |B| = 1500, |A \cap B| \in \{0, 10, \ldots, 1000\}$

red line = $\cos(A, B) = |A \cap B| / \sqrt{|A| \cdot |B|}$

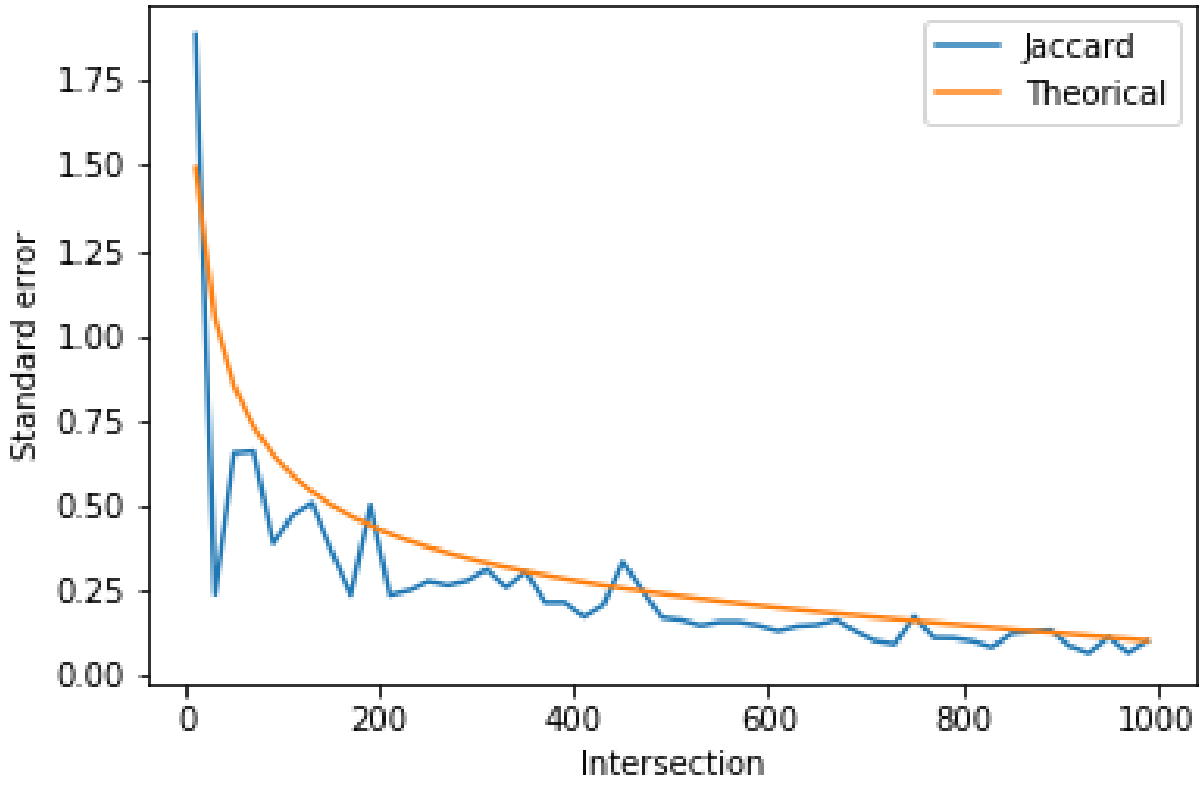
A few simulations



$|A| = 1000, |B| = 1500, |A \cap B| \in \{0, 10, \ldots, 1000\}$

red line = $\mathrm{Corr}(A, B) = |A \cap B|^2/(|A| \cdot |B|)$

A few simulations

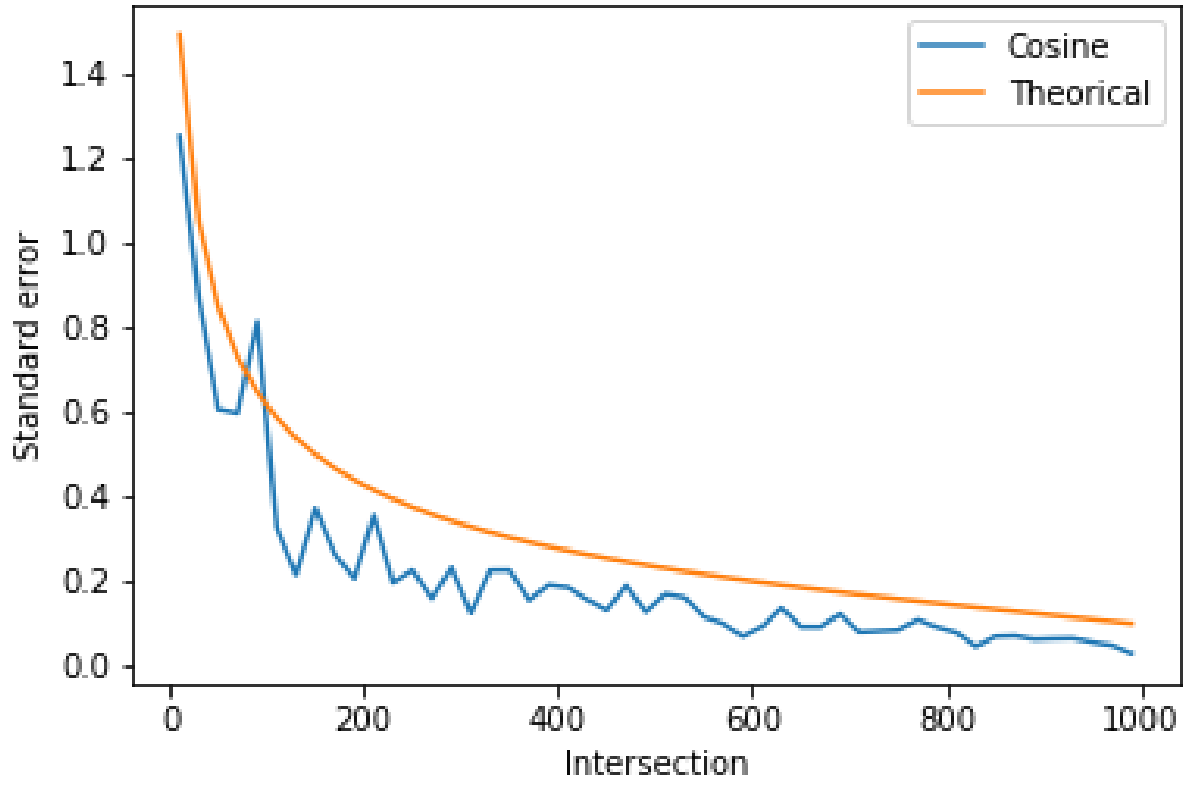


$|A| = 1000, |B| = 1500, |A \cap B| \in \{0, 10, \ldots, 1000\}$

blue line = sample standard error for $J(A, B)$

red line = $\text{SE}_n\,[\hat{\sigma}] = \sqrt{\mathbb{V}\{\hat{\sigma}\}}/\sigma$

A few simulations



$|A| = 1000, |B| = 1500, |A \cap B| \in \{0, 10, \ldots, 1000\}$

blue line = sample standard error for $\cos(A, B)$

red line = $\mathrm{SE}_n[\hat{\sigma}] = \sqrt{\mathbb{V}\{\hat{\sigma}\}}/\sigma$

Some final remarks

Experiments suggest that the standar error of all similarity estimators $\hat{\sigma}$ behave in all cases as

$$\mathrm{SE}_n\left[\hat{\sigma}\right] = \frac{\sqrt{\mathbb{V}\left\{\hat{\sigma}\right\}}}{\mathbb{E}\left\{\hat{\sigma}\right\}} \sim \frac{1}{\sqrt{\sigma \cdot |S|}},$$

not only for Jaccard.
N.B. We can make $|S| \to \infty$ as $|A|, |B| \to \infty$ using affirmative sampling, without commiting to a fixed large value of k, but making the standard error go to 0 (albeit slowly).

Some final remarks

- Prove the conjecture about unbiased estimation of the similarity measures: cosine, correlation, Kulczynski 1
- Prove the conjecture about the standard error for the different studied similarity measures
- Extend our work to other similarity/distance measures: we are aware of almost 80 similarity/distance measures—however many are not interesting in this context as they count also *negative matches* (the number of elements neither in $\mathbb{A}$ nor in $\mathbb{B}$)
- Carry out the analysis of similarity measures for multisets: on-going work, works for Jaccard similarity—however, we need to sample items according to their frequencies, we have not to use distinct sampling