Similarity Search

CSE545 - Spring 2020
Stony Brook University

H. Andrew Schwartz
Big Data Analytics, The Class

**Goal:** Generalizations
A model or summarization of the data.

**Data Frameworks**
- Hadoop File System
- Spark
- MapReduce
- Streaming
- Tensorflow

**Algorithms and Analyses**
- Similarity Search
- Hypothesis Testing
- Link Analysis
- Recommendation Systems
- Deep Learning
Finding Similar Items

(http://blog.soton.ac.uk/hive/2012/05/10/recommendation-system-of-hive/)

(http://www.datacommunitydc.org/blog/2013/08/entity-resolution-for-big-data)
Finding Similar Items: Topics

- Shingling
- Minhashing
- Locality-sensitive hashing
- Distance Metrics
Document Similarity

Challenge: How to represent the document in a way that can be efficiently encoded and compared?
Shingles

**Goal:** Convert documents to sets
**Shingles**

**Goal:** Convert documents to sets

k-shingles (aka “character n-grams”)  
- sequence of $k$ characters

E.g. $k=2$ \( \text{doc} = "abcdabd" \)
\( \text{singles(doc, 2)} = \{ab, bc, cd, da, bd\} \)
Shingles

**Goal:** Convert documents to sets

k-shingles (aka “character n-grams”) - sequence of $k$ characters

E.g. $k=2$ doc=“abcdabcd”
singles(doc, 2) = {ab, bc, cd, da, bd}

- Similar documents have many common shingles
- Changing words or order has minimal effect.
- In practice use $5 < k < 10$
Shingles

**Goal:** Convert documents to sets

- Large enough that any given shingle appearing in a document is highly unlikely (e.g. < 0.1% chance)
- Can hash large shingles to smaller (e.g. 9-shingles into 4 bytes)

*Can also use words (aka n-grams).*

- Similar documents have many common shingles
- Changing words or order has minimal effect.
- **In practice use** $5 < k < 10$
Shingles

**Problem:** Even if hashing, sets of shingles are large (e.g. 4 bytes => 4x the size of the document).
Minhashing

**Goal:** Convert sets to shorter ids, signatures
Minhashing

**Goal:** Convert sets to shorter ids, “signatures”

<table>
<thead>
<tr>
<th>Element</th>
<th>$S_1$</th>
<th>$S_2$</th>
<th>$S_3$</th>
<th>$S_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>b</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>c</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>d</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>e</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

(Leskovec at al., 2014; http://www.mmds.org/)

**Jaccard Similarity:**

\[
sim(S_1, S_2) = \frac{S_1 \cap S_2}{S_1 \cup S_2}
\]

often very sparse! (lots of zeros)
### Minhashing

**Characteristic Matrix:**

<table>
<thead>
<tr>
<th></th>
<th>$S_1$</th>
<th>$S_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ab</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>bc</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>de</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ah</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ha</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ed</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ca</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

**Jaccard Similarity:**

$$sim(S_1, S_2) = \frac{S_1 \cap S_2}{S_1 \cup S_2}$$
Minhashing

Characteristic Matrix:

<table>
<thead>
<tr>
<th></th>
<th>$S_1$</th>
<th>$S_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ab</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>bc</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>de</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ah</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ha</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ed</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ca</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Jaccard Similarity:

$$sim(S_1, S_2) = \frac{S_1 \cap S_2}{S_1 \cup S_2}$$
Minhashing

Characteristic Matrix:

<table>
<thead>
<tr>
<th></th>
<th>$S_1$</th>
<th>$S_2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ab</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>bc</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>de</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ah</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ha</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>ed</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>ca</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Jaccard Similarity:

$$sim(S_1, S_2) = \frac{S_1 \cap S_2}{S_1 \cup S_2}$$

$$sim(S_1, S_2) = \frac{3}{6}$$

# both have / # at least one has
Minhashing

**Problem:** Even if hashing shingle contents, sets of shingles are large.

E.g. 4 byte integer per shingle: assume all unique shingles,

=> 4x the size of the document

(since there are as many shingles as characters and 1 byte per char).
## Minhashing

**Goal:** Convert sets to shorter ids, “signatures”

**Characteristic Matrix:** $X$

<table>
<thead>
<tr>
<th></th>
<th>$S_1$</th>
<th>$S_2$</th>
<th>$S_3$</th>
<th>$S_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ab</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>bc</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>de</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ah</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ha</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ed</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ca</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

(Leskovec et al., 2014; http://www.mmds.org/)
Minhashing

**Goal:** Convert sets to shorter ids, “signatures”

**Characteristic Matrix:** $X$

<table>
<thead>
<tr>
<th></th>
<th>$S_1$</th>
<th>$S_2$</th>
<th>$S_3$</th>
<th>$S_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ab</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>bc</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>de</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ah</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ha</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ed</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ca</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

**Approximate Approach:**

1) Instead of keeping whole characteristic matrix, just keep first row where 1 is encountered.

2) Shuffle and repeat to get a “signature” for each set.

(Leskovec et al., 2014; http://www.mmds.org/)
Minhashing

**Goal:** Convert sets to shorter ids, “signatures”

**Approximate Approach:**
1) Instead of keeping whole characteristic matrix, just keep first row where 1 is encountered.

2) Shuffle and repeat to get a “signature” for each set.

<table>
<thead>
<tr>
<th></th>
<th>$S_1$</th>
<th>$S_2$</th>
<th>$S_3$</th>
<th>$S_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ab</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>bc</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>de</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ah</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ha</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ed</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ca</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

(Leskovec at al., 2014; http://www.mmds.org/)
Minhashing

Goal: Convert sets to shorter ids, “signatures”

Approximate Approach:
1) Instead of keeping whole characteristic matrix, just keep first row where 1 is encountered.

2) Shuffle and repeat to get a “signature”.

(Leskovec at al., 2014; http://www.mmids.org/)
**Minhashing**

**Goal:** Convert sets to shorter ids, “signatures”

---

**Characteristic Matrix: X**

<table>
<thead>
<tr>
<th></th>
<th>S₁</th>
<th>S₂</th>
<th>S₃</th>
<th>S₄</th>
</tr>
</thead>
<tbody>
<tr>
<td>ab</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>bc</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>de</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ah</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ha</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ed</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ca</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

**Approximate Approach:**

1) Instead of keeping whole characteristic matrix, just keep first row where 1 is encountered.

2) Shuffle and repeat to get a “signature”.

---

<table>
<thead>
<tr>
<th>signatures</th>
</tr>
</thead>
<tbody>
<tr>
<td>S₁</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>ah</td>
</tr>
<tr>
<td>ca</td>
</tr>
<tr>
<td>ed</td>
</tr>
<tr>
<td>de</td>
</tr>
<tr>
<td>ab</td>
</tr>
<tr>
<td>bc</td>
</tr>
</tbody>
</table>

(Leskovec et al., 2014; http://www.mmds.org/)
**Minhashing**

**Goal:** Convert sets to shorter ids, “signatures”

**Characteristic Matrix:** \( X \)

<table>
<thead>
<tr>
<th></th>
<th>( S_1 )</th>
<th>( S_2 )</th>
<th>( S_3 )</th>
<th>( S_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>ab</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>bc</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>de</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ah</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ha</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ed</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ca</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

**Approximate Approach:**
1) Instead of keeping whole characteristic matrix, just keep first row where 1 is encountered.

2) Shuffle and repeat to get a “signature” for each set.

**Idea:** We don’t need to actually shuffle. We can just permute row ids.

(Leskovec et al., 2014; http://www.mmds.org/)
Minhashing

Characteristics Matrix:

<table>
<thead>
<tr>
<th></th>
<th>$S_1$</th>
<th>$S_2$</th>
<th>$S_3$</th>
<th>$S_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ab</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>bc</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>de</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ah</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ha</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ed</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ca</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Minhash function: $h$

- Based on permutation of rows in the characteristic matrix, $h$ maps sets to first row where set appears.

(Leskovec et al., 2014; http://www.mmds.org/)
Minhashing

Characteristic Matrix:

<table>
<thead>
<tr>
<th></th>
<th>$S_1$</th>
<th>$S_2$</th>
<th>$S_3$</th>
<th>$S_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ab</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>bc</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>de</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ah</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ha</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>ed</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>ca</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Minhash function: $h$
- Based on permutation of rows in the characteristic matrix, $h$ maps sets to first row where set appears.

(Leskovec at al., 2014; http://www.mmds.org/)
**Minhashing**

Minhash function: $h$
- Based on permutation of rows in the characteristic matrix, $h$ maps sets to first row where set appears.

<table>
<thead>
<tr>
<th></th>
<th>$S_1$</th>
<th>$S_2$</th>
<th>$S_3$</th>
<th>$S_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>ab</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>bc</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>de</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>ah</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>ha</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>ed</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>ca</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Permuted order:
- 1 ha
- 2 ed
- 3 ab
- 4 bc
- 5 ca
- 6 ah
- 7 de

(Leskovec at al., 2014; http://www.mmds.org/)
Minhashing

Minhash function: \( h \)
- Based on permutation of rows in the characteristic matrix, \( h \) maps sets to first row where set appears.

<table>
<thead>
<tr>
<th></th>
<th>( S_1 )</th>
<th>( S_2 )</th>
<th>( S_3 )</th>
<th>( S_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>ab</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>bc</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>de</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>ah</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>ha</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>ed</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>ca</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>permuted order</th>
<th>1</th>
<th>ha</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>ed</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>ab</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>bc</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>ca</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>ah</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>de</td>
<td></td>
</tr>
</tbody>
</table>

\( h(S_1) = ed \)  # permuted row 2
\( h(S_2) = ha \)  # permuted row 1
\( h(S_3) = \)

(Leskovec at al., 2014; http://www.mmds.org/)
Minhashing

Minhash function: \( h \)
- Based on permutation of rows in the characteristic matrix, \( h \) maps sets to first row where set appears.

<table>
<thead>
<tr>
<th></th>
<th>( S_1 )</th>
<th>( S_2 )</th>
<th>( S_3 )</th>
<th>( S_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>ab</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>bc</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>de</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>ah</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>ha</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>ed</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>ca</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

 permuted order

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>ha</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>ed</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>ab</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>bc</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>ca</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>ah</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>de</td>
<td></td>
</tr>
</tbody>
</table>

\( h(S_1) = ed \)  # permuted row 2
\( h(S_2) = ha \)  # permuted row 1
\( h(S_3) = ed \)  # permuted row 2
\( h(S_4) = \)
Minhashing

Minhash function: \( h \)
- Based on permutation of rows in the characteristic matrix, \( h \) maps sets to first row where set appears.

<table>
<thead>
<tr>
<th></th>
<th>( S_1 )</th>
<th>( S_2 )</th>
<th>( S_3 )</th>
<th>( S_4 )</th>
<th>permuted order</th>
</tr>
</thead>
<tbody>
<tr>
<td>3 ab</td>
<td>1 0 1 0</td>
<td>1 ha</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4 bc</td>
<td>1 0 0 1</td>
<td>2 ed</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7 de</td>
<td>0 1 0 1</td>
<td>3 ab</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6 ah</td>
<td>0 1 0 1</td>
<td>4 bc</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 ha</td>
<td>0 1 0 1</td>
<td>5 ca</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2 ed</td>
<td>1 0 1 0</td>
<td>6 ah</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5 ca</td>
<td>1 0 1 0</td>
<td>7 de</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\( h(S_1) = \text{ed} \) \#permuted row 2
\( h(S_2) = \text{ha} \) \#permuted row 1
\( h(S_3) = \text{ed} \) \#permuted row 2
\( h(S_4) = \text{ha} \) \#permuted row 1

(Leskovec at al., 2014; http://www.mmds.org/)
Minhashing

Minhash function: \( h \)
- Based on permutation of rows in the characteristic matrix, \( h \) maps sets to rows.

Signature matrix: \( M \)
- Record first row where each set had a 1 in the given permutation

<table>
<thead>
<tr>
<th></th>
<th>( S_1 )</th>
<th>( S_2 )</th>
<th>( S_3 )</th>
<th>( S_4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>ab</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>bc</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>de</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>ah</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>ha</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>ed</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>ca</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

\( h_1(S_1) = \text{ed} \)  \#permuted row 2
\( h_1(S_2) = \text{ha} \)  \#permuted row 1
\( h_1(S_3) = \text{ed} \)  \#permuted row 2
\( h_1(S_4) = \text{ha} \)  \#permuted row 1

(Leskovec at al., 2014; http://www.mmds.org/)
Minhashing

Characteristic Matrix:

<table>
<thead>
<tr>
<th></th>
<th>$S_1$</th>
<th>$S_2$</th>
<th>$S_3$</th>
<th>$S_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>ab</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>bc</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>de</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>ah</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>ha</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>ed</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>ca</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Minhash function: $h$
- Based on permutation of rows in the characteristic matrix, $h$ maps sets to rows.

Signature matrix: $M$
- Record first row where each set had a 1 in the given permutation

```
<table>
<thead>
<tr>
<th></th>
<th>$S_1$</th>
<th>$S_2$</th>
<th>$S_3$</th>
<th>$S_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h_1$</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>
```

$\text{Permutation}: h_1(S_1) = \text{ed} \#\text{permuted row}

$\text{Permutation}: h_1(S_2) = \text{ha} \#\text{permuted row}$

$\text{Permutation}: h_1(S_3) = \text{ed} \#\text{permuted row}$

(Leskovec at al., 2014; http://www.mmids.org/)
Minhashing

Characteristic Matrix:

<table>
<thead>
<tr>
<th></th>
<th>$S_1$</th>
<th>$S_2$</th>
<th>$S_3$</th>
<th>$S_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>ab</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>bc</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>de</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>ah</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>ha</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>ed</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>ca</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
</tbody>
</table>

Minhash function: $h$

- Based on permutation of rows in the characteristic matrix, $h$ maps sets to rows.

Signature matrix: $M$

- Record first row where each set had a 1 in the given permutation

\[
\begin{array}{cccc}
& S_1 & S_2 & S_3 & S_4 \\
\hline
h_1 & 2 & 1 & 2 & 1 \\
\end{array}
\]

$h_1(S_1) = ed$ #permuted row

$h_1(S_2) = ha$ #permuted row

$h_1(S_3) = ed$ #permuted row

(Leskovec at al., 2014; http://www.mmids.org/)
Minhashing

Characteristic Matrix:

<table>
<thead>
<tr>
<th></th>
<th>$S_1$</th>
<th>$S_2$</th>
<th>$S_3$</th>
<th>$S_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>3</td>
<td>ab</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>bc</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>7</td>
<td>de</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>ah</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>ha</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>ed</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>ca</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Minhash function: $h$
- Based on permutation of rows in the characteristic matrix, $h$ maps sets to rows.

Signature matrix: $M$
- Record first row where each set had a 1 in the given permutation

<table>
<thead>
<tr>
<th></th>
<th>$S_1$</th>
<th>$S_2$</th>
<th>$S_3$</th>
<th>$S_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h_1$</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>$h_2$</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Leskovec at al., 2014; http://www.mmds.org/)
Minhashing

Characteristic Matrix:

<table>
<thead>
<tr>
<th></th>
<th>(S_1)</th>
<th>(S_2)</th>
<th>(S_3)</th>
<th>(S_4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>3</td>
<td>ab</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>bc</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>7</td>
<td>de</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
<td>ah</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
<td>ha</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>ed</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>ca</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Minhash function: \(h\)
- Based on permutation of rows in the characteristic matrix, \(h\) maps sets to rows.

Signature matrix: \(M\)
- Record first row where each set had a 1 in the given permutation

<table>
<thead>
<tr>
<th></th>
<th>(S_1)</th>
<th>(S_2)</th>
<th>(S_3)</th>
<th>(S_4)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(h_1)</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>(h_2)</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>

(Leskovec at al., 2014; http://www.mmds.org/)
Minhashing

Minhash function: $h$
- Based on permutation of rows in the characteristic matrix, $h$ maps sets to rows.

Signature matrix: $M$
- Record first row where each set had a 1 in the given permutation

<table>
<thead>
<tr>
<th>1</th>
<th>4</th>
<th>3</th>
<th>1</th>
<th>0</th>
<th>0</th>
<th>0</th>
<th>0</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>2</td>
<td>4</td>
<td>ab</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>7</td>
<td>bc</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>6</td>
<td>de</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>1</td>
<td>ah</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>2</td>
<td>ha</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>5</td>
<td>ed</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>$S_1$</th>
<th>$S_2$</th>
<th>$S_3$</th>
<th>$S_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h_1$</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>$h_2$</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>$h_3$</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>

(Leskovec at al., 2014; http://www.mmds.org/)
Minhashing

Characteristic Matrix:

<table>
<thead>
<tr>
<th></th>
<th>$S_1$</th>
<th>$S_2$</th>
<th>$S_3$</th>
<th>$S_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>3</td>
<td>ab</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>4</td>
<td>bc</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>7</td>
<td>de</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>6</td>
<td>ah</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>1</td>
<td>ha</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>2</td>
<td>ed</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>5</td>
<td>ca</td>
<td>1</td>
</tr>
</tbody>
</table>

Minhash function: $h$
- Based on permutation of rows in the characteristic matrix, $h$ maps sets to rows.

Signature matrix: $M$
- Record first row where each set had a 1 in the given permutation

<table>
<thead>
<tr>
<th></th>
<th>$S_1$</th>
<th>$S_2$</th>
<th>$S_3$</th>
<th>$S_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h_1$</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>$h_2$</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>$h_3$</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
</tbody>
</table>

(Leskovec at al., 2014; http://www.mmds.org/)
Minhashing

Characteristic Matrix:

<table>
<thead>
<tr>
<th></th>
<th>$S_1$</th>
<th>$S_2$</th>
<th>$S_3$</th>
<th>$S_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>3</td>
<td>ab</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>4</td>
<td>bc</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>7</td>
<td>de</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>6</td>
<td>ah</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>1</td>
<td>ha</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>2</td>
<td>ed</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>5</td>
<td>ca</td>
<td>1</td>
</tr>
</tbody>
</table>

Minhash function: $h$
- Based on permutation of rows in the characteristic matrix, $h$ maps sets to rows.

Signature matrix: $M$
- Record first row where each set had a 1 in the given permutation

<table>
<thead>
<tr>
<th></th>
<th>$S_1$</th>
<th>$S_2$</th>
<th>$S_3$</th>
<th>$S_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$h_1$</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>$h_2$</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>$h_3$</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(Leskovec et al., 2014; http://www.mmds.org/)
**Minhashing**

**Characteristic Matrix:**

<table>
<thead>
<tr>
<th></th>
<th>$S_1$</th>
<th>$S_2$</th>
<th>$S_3$</th>
<th>$S_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>3</td>
<td>ab</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>4</td>
<td>bc</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>7</td>
<td>de</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>6</td>
<td>ah</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>1</td>
<td>ha</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>2</td>
<td>ed</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>5</td>
<td>ca</td>
<td>1</td>
</tr>
</tbody>
</table>

**Property of signature matrix:**

The probability for any $h_i$ (i.e. any row), that $h_i(S_1) = h_i(S_2)$ is the same as Sim($S_1$, $S_2$).

(Leskovec at al., 2014; http://www.mmds.org/)
Minhashing

Property of signature matrix:
The probability for any $h_i$ (i.e. any row), that $h_i(S_1) = h_i(S_2)$ is the same as $\text{Sim}(S_1, S_2)$

Thus, similarity of signatures $S_1, S_2$ is the fraction of minhash functions (i.e. rows) in which they agree.

(Leskovec at al., 2014; http://www.mmds.org/)

<table>
<thead>
<tr>
<th></th>
<th>$S_1$</th>
<th>$S_2$</th>
<th>$S_3$</th>
<th>$S_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>3</td>
<td>ab</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>4</td>
<td>bc</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>7</td>
<td>de</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>6</td>
<td>ah</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>1</td>
<td>ha</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>2</td>
<td>ed</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>5</td>
<td>ca</td>
<td>1</td>
</tr>
<tr>
<td>$h_1$</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>$h_2$</td>
<td></td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>$h_3$</td>
<td></td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Minhashing

Property of signature matrix:
The probability for any \( h_i \) (i.e. any row), that \( h_i(S_1) = h_i(S_2) \) is the same as \( \text{Sim}(S_1, S_2) \)

Thus, similarity of signatures \( S_1, S_2 \) is the fraction of minhash functions (i.e. rows) in which they agree.

Estimate with a random sample of permutations (i.e. \(~100\))

(Leskovec at al., 2014; http://www.mmds.org/)
**Minhashing**

**Property of signature matrix:**
The probability for any $h_i$ (i.e. any row), that $h_i(S_1) = h_i(S_2)$ is the same as Sim($S_1$, $S_2$)

Thus, similarity of signatures $S_1$, $S_2$ is the fraction of minhash functions (i.e. rows) in which they agree.

---

**Characteristic Matrix:**

<table>
<thead>
<tr>
<th></th>
<th>$S_1$</th>
<th>$S_2$</th>
<th>$S_3$</th>
<th>$S_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>3</td>
<td>ab</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>1</td>
<td>ha</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>2</td>
<td>ed</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>5</td>
<td>ca</td>
<td>1</td>
</tr>
</tbody>
</table>

Estimated Sim($S_1$, $S_3$) = agree / all = 2/3

(Leskovec at al., 2014; http://www.mmds.org/)
Minhashing

Property of signature matrix:
The probability for any $h_i$ (i.e. any row), that $h_i(S_1) = h_i(S_2)$ is the same as $\text{Sim}(S_1, S_2)$.

Thus, similarity of signatures $S_1$, $S_2$ is the fraction of minhash functions (i.e. rows) in which they agree.

Estimated $\text{Sim}(S_1, S_3) =$ agree / all $= 2/3$

Real $\text{Sim}(S_1, S_3) =$ Type a / (a + b + c) $= 3/4$

(Leskovec at al., 2014; http://www.mmds.org/)
Minhashing

Characteristic Matrix:

<table>
<thead>
<tr>
<th></th>
<th>$S_1$</th>
<th>$S_2$</th>
<th>$S_3$</th>
<th>$S_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>3</td>
<td>ab</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>4</td>
<td>bc</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>7</td>
<td>de</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>6</td>
<td>ah</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>1</td>
<td>ha</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>2</td>
<td>ed</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>5</td>
<td>ca</td>
<td>1</td>
</tr>
</tbody>
</table>

Property of signature matrix:
The probability for any $h_i$ (i.e. any row), that $h_i(S_1) = h_i(S_2)$ is the same as Sim($S_1$, $S_2$)

Thus, similarity of signatures $S_1$, $S_2$ is the fraction of minhash functions (i.e. rows) in which they agree.

Estimated Sim($S_1$, $S_3$) = agree / all = 2/3

Real Sim($S_1$, $S_3$) = Type a / (a + b + c) = 3/4

Try Sim($S_2$, $S_4$) and Sim($S_1$, $S_2$)

(Leskovec at al., 2014; http://www.mmds.org/)
Minhashing

**Characteristic Matrix:**

| 1 4 3 | 1 0 1 0 |
|---|---|---|---|
| 3 2 4 | 1 0 0 1 |
| 7 1 7 | 0 1 0 1 |
| 6 3 6 | 0 1 0 1 |
| 2 6 1 | 0 1 0 1 |
| 5 7 2 | 1 0 1 0 |
| 4 5 5 | 1 0 1 0 |

**Error Bound?**

Estimated \( \text{Sim}(S_1, S_3) = \frac{2}{3} \)  
Real \( \text{Sim}(S_1, S_3) = \frac{3}{4} \)  
Type a / \( (a + b + c) = \frac{3}{4} \)  
Try \( \text{Sim}(S_2, S_4) \) and \( \text{Sim}(S_1, S_2) \)

(Leskovec at al., 2014; http://www.mmds.org/)
Minhashing

Characteristic Matrix:

<table>
<thead>
<tr>
<th></th>
<th>$S_1$</th>
<th>$S_2$</th>
<th>$S_3$</th>
<th>$S_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>3</td>
<td>ab</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>1</td>
<td>ha</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>4</td>
<td>bc</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>6</td>
<td>ah</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>1</td>
<td>de</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>2</td>
<td>ed</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>5</td>
<td>ca</td>
<td>1</td>
</tr>
</tbody>
</table>

Error Bound?
Expect error: $O(1/\sqrt{k})$ (k hashes)

Why? Each row is a random observation of 1 or 0 (match or not) with $P(match=1) = Sim(S1, S2)$.

Estimated $Sim(S_1, S_3) = agree / all = 2/3$

Real $Sim(S_1, S_3) = Type a / (a + b + c) = 3/4$

Try $Sim(S_2, S_4)$ and $Sim(S_1, S_2)$

(Leskovec at. al., 2014; http://www.mmds.org/)
Minhashing

Characteristic Matrix:

<table>
<thead>
<tr>
<th></th>
<th>S₁</th>
<th>S₂</th>
<th>S₃</th>
<th>S₄</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>3</td>
<td>ab</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>4</td>
<td>bc</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>7</td>
<td>de</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>6</td>
<td>ah</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>1</td>
<td>ha</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>2</td>
<td>ed</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>5</td>
<td>ca</td>
<td>1</td>
</tr>
</tbody>
</table>

Error Bound?
Expect error: $O(1/\sqrt{k})$ (k hashes)

Why? Each row is a random observation of 1 or 0 (match or not) with $P(match=1) = Sim(S₁, S₂)$.

N = k observations
Standard deviation (std)? < 1 (worst case is 0.5)

Estimated $Sim(S₁, S₃) = agree / all = 2/3$

Real $Sim(S₁, S₃) = Type a / (a + b + c) = 3/4$

Try $Sim(S₂, S₄)$ and $Sim(S₁, S₂)$

(Leskovec at al., 2014; http://www.mmds.org/)
**Minhashing**

### Characteristic Matrix:

<table>
<thead>
<tr>
<th></th>
<th>$S_1$</th>
<th>$S_2$</th>
<th>$S_3$</th>
<th>$S_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>3</td>
<td>ab</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>4</td>
<td>bc</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>7</td>
<td>de</td>
<td>0</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>6</td>
<td>ah</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>6</td>
<td>1</td>
<td>ha</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>2</td>
<td>ed</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>5</td>
<td>ca</td>
<td>1</td>
</tr>
</tbody>
</table>

### Error Bound?

**Expect error:** $O(1/\sqrt{k})$ (k hashes)

**Why?** Each row is a random observation of 1 or 0 (match or not) with $P(\text{match}=1) = \text{Sim}(S_1, S_2)$.

**N = k observations**

**Standard deviation (std)?** $< 1$ (worst case is 0.5)

**Standard Error of Mean = std/\sqrt{N}**

### Estimated Sim($S_1, S_3$) =

agree / all = 2/3

### Real Sim($S_1, S_3$) =

Type a / (a + b + c) = 3/4

### Try Sim($S_2, S_4$) and Sim($S_1, S_2$)

(Leskovec at al., 2014; http://www.mmds.org/)
Minhashing

In Practice
Problem:
- Can’t reasonably do permutations (huge space)
- Can’t randomly grab rows according to an order (random disk seeks = slow!)
Minhashing

In Practice
Problem:
- Can’t reasonably do permutations (huge space)
- Can’t randomly grab rows according to an order (random disk seeks = slow!)

Solution: Use “random” hash functions.
- Setup:
  - Pick ~100 hash functions, hashes
  - Store $M[i][s] =$ a potential minimum $h_i(r)$
    
    #initialized to infinity (num hashes x num sets)
Minhashing

Solution: Use “random” hash functions.

Setup:

\[
\text{hashes} = [\text{getHfunc}(i) \text{ for } i \in \text{rand}(1, \text{num}=100)]
\]

#100 hash functions, seeded random

for i in hashes: for s in sets:

\[
\text{Sig}[i][s] = \text{np.inf} \quad \#\text{represents a potential minimum } h_i(r) \; \text{initially infinity}
\]
Minhashing

Solution: Use “random” hash functions.

Setup:

hashes = [getHfunc(i) for i in rand(1, num=100)]

#100 hash functions, seeded random

for i in hashes: for s in sets:
    Sig[i][s] = np.inf #represents a potential minimum $h_i(r)$; initially infinity

Algorithm (“efficient minhashing”):

for r in rows of cm: #cm is characteristic matrix
    compute $h_i(r)$ for all i in hashes #precompute 100 values

for each set s in sets: #columns of cm
    if cm[r][s] == 1:
        for i in hashes: #check which hash produces smallest value
            if $h_i(r) <$ Sig[i][s]: Sig[i][s] = $h_i(r)$
Minhashing

Solution: Use “random” hash functions.

Setup:

    hashes = [getHfunc(i) for i in rand(1, num=100)]

    #100 hash functions, seeded random

    for i in hashes: for s in sets:

        Sig[i][s] = np.inf #represents a potential minimum \( h_i(r) \); initially infinity

Algorithm ("efficient minhashing") without charact matrix:

    for feat in shins: #shins is all unique shingles

        compute \( h_i(\text{feat}) \) for all i in hashes #precompute 100 values

    for each set s in sets: #sets is list of shingle sets

        if feat in s:

            for i in hashes: #check which hash produces smallest value

                if \( h_i(\text{feat}) < \text{Sig}[i][s_{id}] \): \( \text{Sig}[i][s_{id}] = h_i(\text{feat}) \)
Minhashing

**Problem:** Even if hashing, sets of shingles are large (e.g. 4 bytes => 4x the size of the document).
Minhashing

**Problem:** Even if hashing, sets of shingles are large (e.g. 4 bytes => 4x the size of the document).

**New Problem:** Even if the size of signatures are small, it can be computationally expensive to find similar pairs.

E.g. 1m documents; 1,000,000 choose 2 = 500,000,000,000,000 pairs!
Minhashing

**Problem:** Even if hashing, sets of shingles are large (e.g. 4 bytes => 4x the size of the document).

**New Problem:** Even if the size of signatures are small, it can be computationally expensive to find similar pairs.

E.g. 1m documents; 1,000,000 choose 2 = 500,000,000,000,000 pairs!

(1m documents isn’t even “big data”)
Document Similarity

Duplicate web pages (useful for ranking

Plagiarism

Cluster News Articles

Anything similar to documents: movie/music/art tastes, product characteristics

COVID-19 Report matching
Locality-Sensitive Hashing

**Goal:** find pairs of minhashes *likely* to be similar (in order to then test more precisely for similarity).

Candidate pairs: pairs of elements to be evaluated for similarity.
Locality-Sensitive Hashing

**Goal:** find pairs of minhashes *likely* to be similar (in order to then test more precisely for similarity).

Candidate pairs: pairs of elements to be evaluated for similarity.

If we wanted the similarity for all pairs of documents, could anything be done?
**Goal:** find pairs of minhashes likely to be similar (in order to then test more precisely for similarity).

Candidate pairs: pairs of elements to be evaluated for similarity.

Approach: Hash multiple times over subsets of data: similar items are likely in the same bucket once.
Locality-Sensitive Hashing

**Goal:** find pairs of minhashes likely to be similar (in order to then test more precisely for similarity).

Candidate pairs: pairs of elements to be evaluated for similarity.

Approach: Hash multiple times over subsets of data: similar items are likely in the same bucket once.

Approach from MinHash: Hash columns of signature matrix 

→ Candidate pairs end up in the same bucket.
Locality-Sensitive Hashing

Step 1: Divide signature matrix into $b$ bands

$\text{Signature matrix } M$

($\text{Leskovec at al., 2014; http://www.mmds.org/}$)
Locality-Sensitive Hashing

Step 1: Divide into $b$ bands

Will come back to:
Can be tuned to catch most true-positives with least false-positives.

Signature matrix $M$

(Leskovec at al., 2014; http://www.mmds.org/)
Locality-Sensitive Hashing

Step 1: Divide into $b$ bands
Step 2: Hash columns within bands (one hash per band)

(Leskovec at al., 2014; http://www.mmds.org/)
Locality-Sensitive Hashing

Step 1: Divide into $b$ bands
Step 2: Hash columns within bands (one hash per band)

Columns 6 and 7 are surely different.

$r$ rows $b$ bands

(Leskovec at al., 2014; http://www.mmds.org/)
Locality-Sensitive Hashing

Step 1: Divide into $b$ bands
Step 2: Hash columns within bands (one hash per band)

Columns 2 and 6 are probably identical (candidate pair)

Columns 6 and 7 are surely different.

Matrix $M$

$r$ rows

$b$ bands

(Leskovec at al., 2014; http://www.mmds.org/)
Locality-Sensitive Hashing

Step 1: Divide into \( b \) bands
Step 2: Hash columns within bands (one hash per band)

Criteria for being candidate pair:
- They end up in same bucket for at least 1 band.

(Leskovec at al., 2014; http://www.mmds.org/)
Locality-Sensitive Hashing

Step 1: Divide into $b$ bands
Step 2: Hash columns within bands (one hash per band)

Columns 2 and 6 are probably identical (candidate pair)
Columns 6 and 7 are surely different.

Matrix $M$

$r$ rows

$b$ bands

Simplification:
There are enough buckets compared to rows per band that columns must be identical in order to hash into same bucket.

Thus, we only need to check if identical within a band.

(Leskovec at al., 2014; http://www.mmds.org/)
Document-Similarity Pipeline

- Shingling
- Minhashing
- Locality-sensitive hashing
Probability of Agreement

- 100,000 documents
- 100 random permutations/hash functions/rows
  => if 4-byte integers then 40Mb to hold signature matrix
  => still 100k choose 2 is a lot (~5 billion)
Probability of Agreement

- 100,000 documents
- 100 random permutations/hash functions/rows
  => if 4byte integers then 40Mb to hold signature matrix
  => still 100k choose 2 is a lot (~5billion)
- 20 bands of 5 rows
- Want 80% Jaccard Similarity ; for any row $p(S_1 == S_2) = .8$
100,000 documents
100 random permutations/hash functions/rows
  => if 4-byte integers then 40Mb to hold signature matrix
  => still 100k choose 2 is a lot (~5 billion)
20 bands of 5 rows
Want 80% Jaccard Similarity; for any row \( p(S_1 == S_2) = .8 \)

\( P(S_1 == S_2 \mid b^{(5)}) \): probability \( S_1 \) and \( S_2 \) agree within a given band
Probability of Agreement

- 100,000 documents
- 100 random permutations/hash functions/rows
  => if 4-byte integers then 40Mb to hold signature matrix
  => still 100k choose 2 is a lot (~5 billion)
- 20 bands of 5 rows
- Want 80% Jaccard Similarity; for any row \( p(S_1 = S_2) = .8 \)

\[
P(S_1 = S_2 | b^{(5)}) \text{: probability S1 and S2 agree within a given band} \]
\[
= 0.8^5 = .328
\]

(Leskovec at al., 2014; http://www.mmds.org/)
Probability of Agreement

- 100,000 documents
- 100 random permutations/hash functions/rows
  => if 4byte integers then 40Mb to hold signature matrix
  => still 100k choose 2 is a lot (~5billion)
- 20 bands of 5 rows
- Want 80% Jaccard Similarity; for any row \( p(S_1 = S_2) = .8 \)

\[
P(S_1 = S_2 \mid b^{(5)})\text{: probability S1 and S2 agree within a given band} = 0.8^5 = .328 \quad \Rightarrow \quad P(S_1 \neq S_2 \mid b) = 1 - .328 = .672
\]

(Leskovec at al., 2014; http://www.mmds.org/)
Probability of Agreement

- 100,000 documents
- 100 random permutations/hash functions/rows
  => if 4-byte integers then 40Mb to hold signature matrix
  => still 100k choose 2 is a lot (~5 billion)
- 20 bands of 5 rows
- Want 80% Jaccard Similarity; for any row \( p(S_1 = S_2) = .8 \)

\[
P(S_1 = S_2 \mid b^{(5)}) \text{: probability S1 and S2 agree within a given band} = 0.8^5 = .328 \quad \Rightarrow \quad P(S_1 != S_2 \mid b) = 1 - .328 = .672
\]

\[
P(S_1 != S_2) \text{: probability S1 and S2 do not agree in any band}
\]

(Leskovec et al., 2014; http://www.mmds.org/)
Probability of Agreement

- 100,000 documents
- 100 random permutations/hash functions/rows
  => if 4byte integers then 40Mb to hold signature matrix
  => still 100k choose 2 is a lot (~5billion)
- 20 bands of 5 rows
- Want 80% Jaccard Similarity; for any row \( P(S_1 = S_2) = 0.8 \)

\[
P(S_1 = S_2 \mid b^{(5)}) \text{: probability S1 and S2 agree within a given band} \\
= 0.8^5 = 0.328 \Rightarrow P(S_1 \neq S_2 \mid b) = 1 - 0.328 = 0.672
\]

\[
P(S_1 \neq S_2) \text{: probability S1 and S2 do not agree in any band} \\
= 0.672^{20} \\approx 0.00035
\]

(Leskovec at al., 2014; http://www.mmds.org/)
Probability of Agreement

- 100,000 documents
- 100 random permutations/hash functions/rows
  => if 4-byte integers then 40Mb to hold signature matrix
  => still 100k choose 2 is a lot (~5 billion)
- 20 bands of 5 rows
- Want 80% Jaccard Similarity; for any row \( p(S_1 == S_2) = .8 \)

\[
P(S_1 == S_2 \mid b): \text{probability } \text{S1 and S2 agree within a given band} \\
= 0.8^5 = .328 \quad \Rightarrow \quad P(S_1 != S_2 \mid b) = 1-.328 = .672
\]

\[
P(S_1 != S_2): \text{probability } \text{S1 and S2 do not agree in any band} \\
=.672^{20} = .00035
\]

What if wanting 40% Jaccard Similarity?
Document-Similarity Pipeline

Shingling → Minhashing → Locality-sensitive hashing