

WHAT IS CLAIMED IS:

1. A method, comprising:
 - receiving, by a processor, a plurality of artifact fingerprints generated from a plurality of artifacts, wherein the plurality of artifact fingerprints are
 - 5 generated by
 - generating shingles from text within each of the plurality of artifacts; and
 - cryptographically hashing the shingles to generate a plurality of artifact fingerprints;
 - 10 storing the artifact fingerprints in a database;
 - receiving, by the processor, at least one query fingerprint from a querying entity;
 - determining, by the processor, whether the received at least one query fingerprint matches any of the artifact fingerprints stored in the database; and
 - 15 outputting an indication of an entity that provided an artifact containing a matched fingerprint to the querying entity without revealing plaintext of the artifact containing the matched fingerprint,
 - wherein the plurality of artifact fingerprints are received from a plurality of different entities,
 - 20 wherein the at least one query fingerprint is generated by:
 - receiving a plaintext query;
 - generating a plurality of query shingles from the plaintext query so that character overlap between adjacent shingles is greater than a character overlap between adjacent shingles of the plurality of artifact fingerprints;
 - 25 cryptographically hashing the plurality of query shingles to form a plurality of query fingerprints.
2. The method of claim 1, wherein the determination of whether the received at least one query fingerprint matches any of the stored artifact
- 30 fingerprints comprises:
 - generating, by the processor, a plurality of query vectors from at least one cryptographically hashed shingle of the query fingerprint;

generating, by the processor, a plurality of artifact vectors from the artifact fingerprints;

performing, by the processor, a cosine distance calculation using the plurality of query vectors and the plurality of artifact vectors to generate a
5 plurality of distance scores; and

selecting, by the processor, a vector of the plurality of artifact vectors having a lowest distance score.

3. The method of claim 2, further comprising:

10 determining, by the processor, a median cosine distance for the plurality of artifact vectors, wherein the median cosine distance is used to set a threshold against which the lowest distance score is compared to determine whether the vector of the plurality of artifact vectors having the lowest distance score is a match of one or more of the query vectors.

15

4. The method of claim 1, wherein the processor and database are maintained by an entity that is independent of the plurality of different entities that provided the plurality of fingerprints and independent of the querying entity and the plurality of artifact fingerprints are generated by each of the plurality of
20 different entities or are generated by the processor that receives the artifacts from the plurality of different entities.

5. The method of claim 1, wherein the processor and database are maintained by the querying entity and the plurality of fingerprints are generated
25 by each of the plurality of different entities or generated by an entity that is independent of the plurality of different entities and independent of the querying entity.

6. The method of claim 1, wherein the shingles generated from the text
30 within each of the plurality of artifacts and the shingles generated from the plaintext query are salted prior to the cryptographic hashing.

7. The method of claim 1, wherein prior to cryptographically hashing the shingles generated from the text within each of the plurality of artifacts, the method further comprises:
- comparing the shingles generated from the text within each of the
 - 5 plurality of artifacts to a stop word list; and
 - removing any shingle of the shingles generated from the text within each of the plurality of artifacts that match a word in the stop word list.
8. The method of claim 1, wherein prior to determining whether the received
- 10 at least one fingerprint matches any of the artifact fingerprints stored in the database, the method comprises:
- receiving a set of artificially common fingerprints; and
 - removing any fingerprints from the stored artifact fingerprints matching any of the fingerprints in the set of artificially common fingerprints.
- 15
9. The method of claim 1, wherein the database storing the artifact fingerprints is a distributed key store.
10. The method of claim 1, further comprising:
- 20 contacting, by the querying entity, the entity that provided the artifact containing the matched fingerprint to obtain a plaintext version of the artifact containing the matched fingerprint.
11. The method of claim 10, further comprising:
- 25 encrypting, by the contacted entity, the artifact containing the matched fingerprint;
- transmitting the encrypted artifact to the querying entity; and
 - decrypting, by the querying entity, the encrypted artifact to obtain the plaintext version of the artifact containing the matched fingerprint.
- 30
12. A system, comprising:
- a memory; and

a processor coupled to the memory, wherein when the processor
 executes computer code stored in the memory, the processor
 receives a plurality of artifact fingerprints generated from a plurality of
 artifacts, wherein the plurality of artifact fingerprints are generated by
 5 generating shingles from text within each of the plurality of
 artifacts; and
 cryptographically hashing the shingles to generate a
 plurality of artifact fingerprints;
 stores the artifact fingerprints in the memory;
 10 receives at least one query fingerprint from a querying entity;
 determines whether the received at least one query fingerprint matches
 any of the artifact fingerprints stored in the database; and
 outputs an indication of an entity that provided an artifact containing a
 matched fingerprint to the querying entity without revealing plaintext of the
 15 artifact containing the matched fingerprint,
 wherein the plurality of artifact fingerprints are received from a plurality of
 different entities,
 wherein the at least one query fingerprint is generated by:
 receiving a plaintext query;
 20 generating a plurality of query shingles from the plaintext query so that
 character overlap between adjacent shingles is greater than a character overlap
 between adjacent shingles of the plurality of artifact fingerprints;
 cryptographically hashing the plurality of query shingles to form a
 plurality of query fingerprints.
 25

13. The system of claim 12, wherein the processor executes the computer
 code for the determination of whether the received at least one query fingerprint
 matches any of the stored artifact fingerprints by:
 generating a plurality of query vectors from at least one cryptographically
 30 hashed shingle of the query fingerprint;
 generating a plurality of artifact vectors from the artifact fingerprints;

performing a cosine distance calculation using the plurality of query vectors and the plurality of artifact vectors to generate a plurality of distance scores; and

5 selecting a vector of the plurality of artifact vectors having a lowest distance score.

14. The system of claim 13, wherein the processor executes the computer code to:

10 determine a median cosine distance for the plurality of artifact vectors, wherein the median cosine distance is used to set a threshold against which the lowest distance score is compared to determine whether the vector of the plurality of artifact vectors having the lowest distance score is a match of one or more of the query vectors.