T-79.514 Special Course on Cryptology

Vulnerabilities in Similarity Search Based Systems

Sami Vaarala
Helsinki University of Technology
`sami.vaarala@iki.fi`

Overview

The discussion is based on: Ali Saman Tosun, Hakan Ferhatosmanoglu: *Vulnerabilities in Similarity Search Based Systems*, CIKM 2002 (Conference on Information and Knowledge Management).

- Database model and problem, reply and score models
- Vulnerabilities in reply and score models
- Attack detection
- Practical experiments (by authors)
- Thoughts

Model and problem

n-dimensional database with valuable, proprietary data (multimedia, CAD/CAM, medical imaging, ...)

- Want to allow similarity search but don't want to reveal data
- Distance function between two elements

Two query models

- *reply model*: return database element closest to query
- *score model*: return score of database element closest to query (distance or a function of distance)

Relation to PPDM

The paper only assumes knowledge of element closest to query (reply model) or score of element closest to query (score model).

Typical privacy-preserving assumption is that Alice may only know the result and anything directly derivable from that (and Bob may know neither the query nor the result). The attacks described work only with such data; thus they apply regardless of the privacy preserving protocol used.

Bob knows query and query result, no effort to hide that (contrast to e.g. Du and Atallah). This is not central though, the focus is on Alice's ability to reconstruct the database.

Security issues considered

Attacker – best strategy to copy whole database or learn contents?

Defender – what can be done to prevent (detect) such attacks?

Relevance – similarity search can make database reconstruction (at least approximate) more feasible

Authors do not define any particular criteria for evaluating the effectiveness of attacks.

Reply model vulnerabilities

General scheme is straightforward

- Assume shortest distance between elements is c, and use an equally spaced grid of queries (e.g. in 2D $c/\sqrt{2}$) so that at most one element in each query point's Voronoi region $\Rightarrow$ Database recovered in full if c correct
- If c is too large, attacker gets a rough characterization of database "shape"

Reply model vulnerabilities

Variations: progressive, adaptive progressive, random, distributed.

Complexity of attack: $O(k^n)$, n dimensions, k depends on c and range of element coordinates.

Score model vulnerabilities

Basic problem

- Given a score, how to increase score and eventually find element?
- Authors assume score equals distance, and distance function is known: $d(x, y) = (\Sigma(x_i - y_i)^2)^{1/2}$

Solution

- Solution $\Rightarrow$ perform multiple queries and solve unknown coordinates from set of equations. Knowledge of the distance function is essential in mounting the attack.

Score model vulnerabilities

Works only if all queries return score for the same element; resulting potential element can be verified (assuming $d(x, y) = 0 \text{ iff } x = y$)

Complexity of attack: $O(nk^n)$, n dimensions, k depends on c and range of element coordinates. (Assuming no “misses” in coordinate reconstruction.)

Attack detection

Attack detection (reply model)

- Detect patterns in queries (difficult: reordering queries, pseudo or true random queries)

- Limit amount of queries per client

- Client identification using query source IP address – this is a bad idea because of private address spaces and Network Address Translation (NAT)

- Distributed attacks – try to use IP address class (A, B, C) to detect requests from the same network (unreliable)

Attack detection

Attack detection (score model)

- “Query similarity function” – sum of coordinates of query point (weak); authors also encode IP: (192.80.25.6, query [2 5 8.2], sum 192080025006015.2)

- Idea: compress multidimensional query into one dimension, “closeness” of two queries can still be determined (although some false positives)

- Replicated databases – databases need to coordinate request handling (seems unrealistic)

Practical experiments (by authors)

Reply model, using random and adaptive progressive

- Data sets – all two-dimensional $\Rightarrow$ little information on practical feasibility
  - $\star$ (lat, long) of road crossings in Maryland
  - $\star$ Stock time series, (random) clustered points
- Results: initially no difference, after refinement adaptive clearly better
- Since number of grid points is $O(k^n)$ for some k dependent on database and c; higher dimensions more interesting.

Practical experiments (by authors)

Score model, investigate how well “similarity function” works, no actual attack

- Stock time series (360 days (dimension), 6500 companies (elements))
- Speech recognition database (618 dimensions, 7800 points)
- Results – query similarity sum threshold should be low, three strike policy used

Thoughts

Practicality

- 10 dimensions, attributes values $[1, 100]$, grid spacing 1 $\Rightarrow$ 100^{10} queries in basic reply model; a billion queries / second means 3170 years, but adaptation not taken into account (incomplete results are also interesting)

Tweaking score model (i.e. score is some function of distance)

- apply ceiling to distance function (nb: query with high distance in result reveals a lot of what is *not* in database!)

- discretize distance – does not seem very useful

Thoughts

How useful is an approximate reconstruction (e.g. reply model with c too large)?

- “Shape” of database is recovered but not its density
- Relevance depends on application

Performance metrics

- Exhaustive search (assuming some c) is always possible, some performance metric would be needed to make useful comparisons or to test feasibility

Conclusions

Ali Saman Tosun, Hakan Ferhatosmanoglu: *Vulnerabilities in Similarity Search Based Systems*, CIKM 2002.

Alice gets reply or score, Bob knows query and query result. Try to prevent database reconstruction from similarity search results.

Experimental results are inconclusive; reply model data was 2-dimensional, score model attacks not directly considered.

If score equals distance and distance function known, score model is not much more secure than reply model. Some alternatives considered.