Anonymization of Spatial Data by Gaussian Skew: Is Re-Identification Possible?

Christopher A. Cassa, M.Eng, Kenneth D. Mandl, M.D. M.P.H.

Children's Hospital Informatics Program at the Harvard-MIT Div of Health Sciences and Tech.

Objective

To evaluate the robustness of a spatial anonymization algorithm for syndromic surveillance data against a triangulation vulnerability attack.

`Background

We have published an anonymization algorithm that takes precise point locations for patients and moves them a randomized distance according to a 2D Gaussian distribution that is inversely adjusted by the underlying population density. Before such algorithms can be integrated into live systems, assurances are needed so that patients cannot be reidentified through systematic vulnerabilities. Here we investigate the ease with which a spatial anonymization algorithm can be compromised by triangulating the original points with multiple repeated data requests.

Obfuscative and cryptographic algorithms may be susceptible to weakening when it is possible for an adversary to produce output from the algorithm according to adversary-provided input. Under this threat model, an adversary could use a syndromic surveillance system to request anonymized patient data from a RHIO or other health network several different times. If the anonymized results are produced each time they are requested, triangulation of original addresses may be possible or the anonymity afforded by the algorithm may be reduced.

Methods

A dataset containing artificially generated geocoded values for 10,000 sample patients was created using a spatial cluster creation tool [1]. Each of the geocoded addresses was anonymized using a Gaussian 2D spatial blur skew that was adjusted for population density [2] a total of 50 times.

With each subsequent anonymization pass, the geocoded datapoints that referred to the same individual address were averaged. After each pass, the distance between the average anonymized point and the original address were calculated.

Results

Reverse identification attempts significantly weakened the anonymity afforded by the spatial dataset. The average distance to the original addresses after one anonymization pass, which represents normal use of an anonymizing algorithm, was 0.69km. After each point was triangulated using the average of fifty anonymization passes, the average distance to the original point in the dataset was reduced to 0.1km.

The average distance to the original address is plotted versus the number of separate anonymization passes in Figure 1. There is a sharp decrease in the average distance to the original address with 10 anonymization passes (0.5km reduction in 10 passes) and thus a sharp decrease in dataset anonymity.

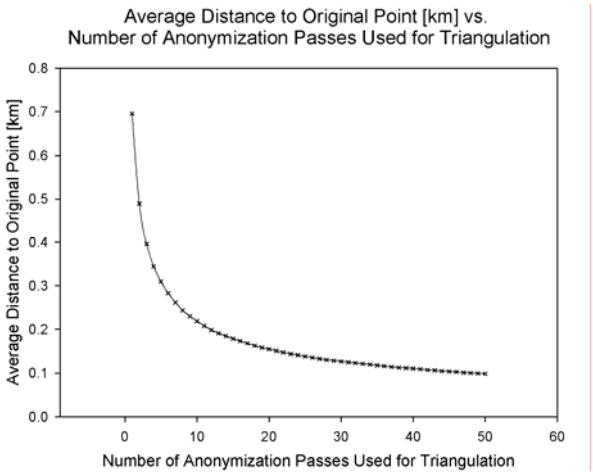

Figure 1 – Average distance to original point [km] vs. number of anonymization passes used for triangulation.

Conclusions

In order to protect privacy when using spatial skew algorithms, the number of distinct anonymization results or passes that represent the same data must be controlled. Limiting the number of passes will avoid reidentification through triangulation. An alternative approach would be for the data provider to maintain and release the same anonymized patient address data for every record requested by a syndromic surveillance system, This approach avoids running the algorthim anew with each request, and introducing the variation that is at the root of the vulnerability.

Another possible intervention might be to swap the addresses in a given dataset so that they are effectively unlinked with any identifiers. An attack of the nature described here relies on linking the anonymized data together using identifiers, demographic or clinical data. This unlinking of spatial data from unique identifiers, however, poses additional challenges for making reverse identification possible.

References

[1] Cassa CA, Iancu K, Olson KL, Mandl KD. A software tool for creating simulated outbreaks to benchmark surveillance systems. *BMC Med Inform Decis Mak.* Jul 14 2005;5(1):22-28.

[2] Cassa CA, Grannis SJ, Overhage JM, Mandl KD. A context-sensitive approach to anonymizing spatial surveillance data: impact on outbreak detection. J Am Med Inform Assoc 2006;13(2):160-5.

Further Information:
Christopher Cassa, cassa@mit.edu, www.chip.org