Fast Graph Scan for Scalable Detection of Arbitrary Connected Clusters

Skyler Speakman, M.S., Daniel B. Neill, Ph.D.

H.J. Heinz III College, Carnegie Mellon University, Pittsburgh, PA 15213

OBJECTIVE

This work presents GraphScan, a spatial scan method for detection of arbitrarily-shaped connected clusters. GraphScan enables efficient, exact computation of the highest-scoring connected clusters, with or without proximity constraints, up to ~100 locations.

BACKGROUND

FlexScan [1] extends Kulldorff's original spatial scan [2] to detect flexibly-shaped clusters consisting of a center location s_c and a connected subset of its $k – 1$ nearest neighbors. Unlike other graph-based scan methods [3], FlexScan finds the highest-scoring connected subgraph, subject to the constraint on neighborhood size k. However, its run time scales exponentially with k, and thus it is computationally infeasible (requiring over 1 week to find the highest-scoring cluster for a single day of data) for $k > 30$.

Linear-time subset scanning [4] can find the most interesting subset of N locations without exhaustively searching over the exponentially many subsets. Many commonly used scan statistics satisfy the LTSS property, which guarantees that the highest scoring subset will consist of the j highest priority locations, for some priority function $G(s_i)$ and j   {1…N}. In this case, only N of the 2^N subsets must be evaluated. For example, in Kulldorff's statistic [2], we can use $G(s_i) = c_i / b_i$, the ratio of observed to expected count.

METHODS

While the unconstrained LTSS method may return a disconnected subset of locations, GraphScan expands on LTSS by only considering the *connected* subsets that have potential for highest score. For a score function that satisfies the LTSS property, we show that if location s_i is contained in the optimal subset S^*, and if removing s_i does not disconnect the subgraph, any neighbor of s_i with higher priority will also be contained in S^*. GraphScan uses this property and the graph structure to reduce the search space, pruning any subgraphs which violate the rule. This efficient search allows GraphScan to abandon proximity constraints and feasibly search over *all* connected subsets of nodes. Alternatively, GraphScan can be used to detect the same proximity-constrained connected clusters as FlexScan, but can scale up to much larger neighborhood sizes (higher values of k).

RESULTS

The GraphScan method was evaluated using Emergency Department data from 91 Allegheny County zip codes. It identified the highest-scoring connected cluster in an average of 125 seconds for each day of data. Figure 1 summarizes the average proximity-constrained run times as a function of the neighborhood size k. GraphScan required less than 1 minute of run time per day of data for all $k \leq 56$.

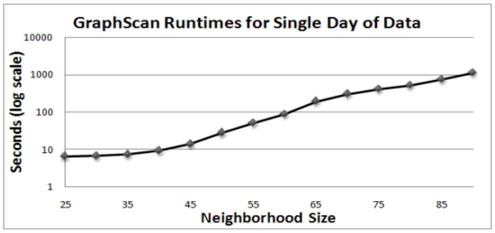

Figure 1 – Run times for GraphScan with proximity constraints.

We compared the detection power and spatial accuracy of GraphScan ($k = 25$ and $k = 50$) to the original Kulldorff scan statistic [2] (circular regions) on elongated semi-synthetic outbreaks injected into the Allegheny ED data. Kulldorff's method detected 76% of injects in an average of 10.0 days, and had average spatial precision of 33% and recall of 32% at the midpoint (day 7) of the outbreak. GraphScan with neighborhood size $k = 25$ detected 95% of injects, had average precision of 37% and recall of 40%, and detected outbreaks an average of 1.9 days earlier. Increasing k to 50 improved detection time by an additional 0.4 days and increased recall to 52%, while slightly reducing precision to 31%.

CONCLUSIONS

Our results demonstrate that GraphScan can improve detection power by accurately identifying irregularly-shaped clusters, and can scale up to much larger neighborhoods than FlexScan (~100 locations). We are currently incorporating fast branch-and-bound algorithms into GraphScan, and anticipate that these improvements will dramatically increase the size of datasets for which it can be feasibly used.

This work was partially supported by NSF grants IIS-0916345, IIS-0911032, and IIS-0325581.

REFERENCES

[1] Tango T, Takahashi K. A flexibly shaped spatial scan statistic for detecting clusters. Intl J Health Geographics, 2005, 4: 11.

[2] Kulldorff M. A spatial scan statistic. Communications in Statistics: Theory and Methods, 1997, 26(6): 1481-1496.

[3] Patil GP, Taillie C. Upper level set scan statistic for detecting arbitrarily shaped hotspots. Envir. Ecol. Stat., 2004, 11:183-197.

[4] Neill DB. Fast and flexible outbreak detection by linear-time subset scanning. Adv Disease Surveillance, 2008, 5:48.

Further Information:
Daniel B. Neill, neill@cs.cmu.edu
http://www.cs.cmu.edu/~neill