SANDIA REPORT

SAND2018-11123
Unlimited Release
Printed October 2018

Adverse Event Prediction Using Graph-Augmented Temporal Analysis: Final Report

Randy C. Brost, Erin E. Carrier, Michelle J. Carroll, Katrina M. Groth,
W. Philip Kegelmeyer,Vitus J. Leung, Hamilton E. Link, Andrew J. Patterson,
Cynthia A. Phillips, Samuel Richter, David Robinson, Andrea Staid,
Diane M.-K. Woodbridge

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by National Technology and Engineering Solutions of Sandia, LLC.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.fedworld.gov
Online ordering: http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2018-11123
Unlimited Release
Printed October 2018

Adverse Event Prediction Using Graph-Augmented Temporal Analysis: Final Report

Randy C. Brost[1], Erin E. Carrier[2], Michelle J. Carroll[1], Katrina M. Groth[3], W. Philip Kegelmeyer[4], Vitus J. Leung[1], Hamilton E. Link[1], Andrew J. Patterson[1], Cynthia A. Phillips[1], Samuel Richter[5], David Robinson[1], Andrea Staid[1], Diane M.-K. Woodbridge[6]

[1]Sandia National Laboratories
P.O. Box 5800
Albuquerque, NM 87185

[2]Department of Computer Science
University of Illinois at Urbana-Champaign
201 N. Goodwin Ave.
Urbana, IL 61802

[3]Department of Mechanical Engineering
0151C Glenn L. Martin Hall
University of Maryland
College Park, MD 20742

[4]Sandia National Laboratories
PO Box 969
Livermore, CA 94551-0969

[5] Missouri University of Science and Technology
205 W 12th St
Rolla, MO 65409

[6]Master of Science in Analytics Program
University of San Francisco
San Francisco, CA 94117

Abstract

This report summarizes the work performed under the Sandia LDRD project "Adverse Event Prediction Using Graph-Augmented Temporal Analysis." The goal of the project was to develop a method for analyzing multiple time-series data streams to identify precursors providing advance warning of the potential occurrence of events of interest. The proposed approach combined temporal analysis of each data stream with reasoning about relationships between data streams using a geospatial-temporal semantic graph. This class of problems is relevant to several important topics of national interest. In the course of this work we developed new temporal analysis techniques, including temporal analysis using Markov Chain Monte Carlo techniques, temporal shift algorithms to refine forecasts, and a version of Ripley's K-function extended to support temporal precursor identification. This report summarizes the project's major accomplishments, and gathers the abstracts and references for the publication submissions and reports that were prepared as part of this work. We then describe work in progress that is not yet ready for publication.

Acknowledgment

This work was funded by the Sandia National Laboratories Laboratory Directed Research and Development program in the Computing and Information Science investment area. We thank Kristina Czuchlewski, John Feddema, Kathy Simonson, and John Wagner for helpful comments and discussions. Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE NA0003525.

Contents

List of Figures

Chapter 1

Summary

1.1 Motivation

In this work we addressed the difficult problem of searching complex data streams for patterns leading up to adverse events. Given a time series of remote sensing and other data, we proposed a method for searching temporal data for precursors leading up to adverse events. Identification of such precursors could aid both forensic event analysis and prediction of a possible future event.

We recognized that for realistic problems there can be an enormous number of possible causal relationships, and so we proposed to augment the temporal analysis with relationship information inferred from a geospatial-temporal semantic graph.

Figure 1.1 shows a schematic view of the proposed approach. At left is a geospatial region with multiple data sources, each of which generates a time series of data. Moving left to right, these data streams are subject to temporal analysis, such as periodicity analysis, outlier identification, and inter-series relationships. This pure temporal analysis identifies candidate precursors, shown with hypothesized correlation weights in the lower right. However, some of these identified correlations may be implausible, based on other information. A graph encoding such information is shown in matrix form on the lower left. These graph relationships reinforce some candidate correlations, while deprecating others.

Writ large, the proposed idea asserts that graph information would allow us to prune the candidate precursors found by temporal analysis, thus exploiting the information content of the graph to reduce computation time and focus output to a relevant subset. Our goal was to implement and test this approach on a number of examples.

Identifying temporal anomalies and relevant correlations is a problem encountered in multiple national security domains, and a solution to this problem could benefit multiple government agencies. Such a result would help analysts find important temporal relationships that are currently difficult to detect, supporting critical decision-making.

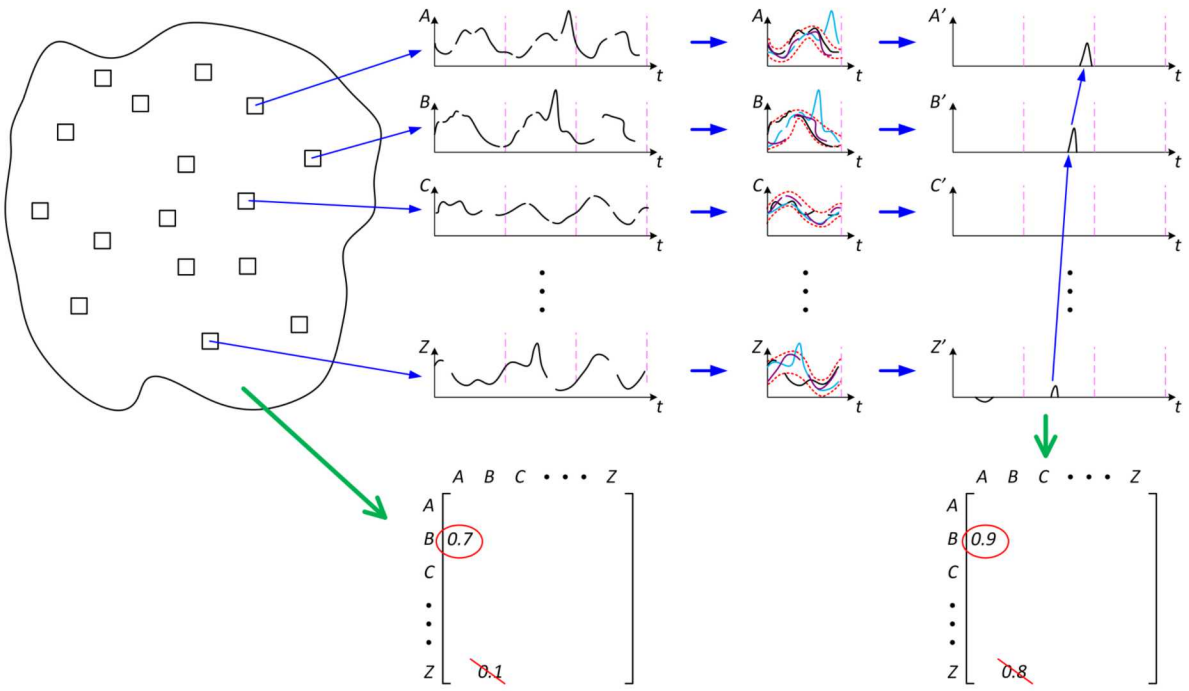

Figure 1.1. Envisioned general approach.

1.2 Survey of Problems and Data

The proposed research embraces a wide-open horizon of potential example problems. Thus our first activity was to survey this range of problems, to understand what classes of problems characterize it, and how they might be grouped according to common properties, common solution approaches, etc. We generated a long list of example candidate problems to consider, and studied their similarities and differences. A sampling of problems considered appears below:

- Site computer network data.
- Dengue / Ebola / Influenza forecasting.
- City police misconduct.
- Philadelphia traffic/crime/event analysis, including bus and taxi operations, etc.
- Trajectories, airport closure events, etc.
- Electric signals in Earth's atmosphere, predicting earthquakes.
- Medical data from hospitals.
- Various national security examples.
- Data describing well injection and seismic events in Oklahoma.
- Analysis of media information provided by GDELT [11].
- Terror microcycles [1].
- Moose-vehicle collisions in rural Maine [5].
- Power grid analysis.

The full list of considered problems is much larger, but these are among the most important suggestions. When considering these lists of problems, several questions naturally arose: What makes a good problem? How do we choose?

During the course of our study, we identified these aspects as characteristics of a problem with desirable properties:

- Relevant.
- Well-defined problem.
- Suitable for publication.
- Amenable to combined temporal/graph analysis.
- Needs graph component, preferably a geospatial graph.
- Multi-modality.
- "Events" are clear.
- Different information sources combine to indicate an event precursor.
- Good data.

Several, but not all, of our candidates met most of the above criteria. The final criterion, "good data," raised another question: What constitutes a good data set, for our purposes? In response we developed the following criteria of a desired data set:

- Available.
- Both spatially and temporally dense.
- Contains needed content. Extraneous content is a plus.
- Includes all key needed data, such as travel patterns.
- Discretization done, if applicable.
- Landcover maps exist, if relevant.
- Ground truth available.

As explained in Section 1.4 below, many data sets fell short of these characteristics, particularly in simultaneous spatial and temporal density, or the availability of ground truth.

Setting aside data concerns, we also considered a range of problems and evaluated suitable solution approaches. We found that a key discriminator that drove the selection of solution approaches was the cardinality of the input data. Input data sets tended to be either continuous, or binary (event-based). Further, it is useful to distinguish whether the data in question are defining the events of interest, or the background or "sensor" data. Grouping data sets according to these criteria yields the problem/solution taxonomy table shown in Figure 1.2.

For this project we are interested in predicting events, which are binary response variables. Thus our attention focused on the right-hand column of the table.

		Following	
		Continuous	Binary
Preceding	Continuous	Granger Causality	Logistical Regression, Poisson Regression
	Binary		K-Function, Transfer Entropy

Figure 1.2. Taxonomy of prediction problems.

1.3 Focus Problems

Based on all of this survey and analysis, we selected four key challenge problems to focus on for the remainder of our work:

1. *Dengue fever outbreak prediction.* Given a data stream of dengue fever cases and related contextual data, con we predict dengue fever outbreaks?
2. *Wind forecast error correction.* Given a data stream of power production from wind power generation facilities and associated forecasts, can we predict forecast errors?
3. *Network precursor analysis.* Given a data stream of network traffic and instances of events, can we identify precursors which warn of impending events?
4. An example national security problem, explained elsewhere.

These problems contain a mixture of continuous and discrete data sets. They also provide different perspectives on the prediction problem, avoiding a too-narrow focus for the project. In addition, in this exploratory project we did not know which approaches would be most successful, and this selection of problems allowed pursuit of different approaches, adding technical diversity and reducing risk. Perhaps most decisively, all of these problems had available data sets.

Our approach and results for these challenge problems are summarized briefly below:

1. *Dengue fever.*
   We obtained a data set listing dengue fever cases and contemporaneous weather and other contextual data from the Dengue Forecasting Project [10]. We first approached this problem by applying Markov Chain Monte Carlo (MCMC) analysis to the entire data set, producing a multivariate model that identified the nominal seasonal variation, and the dengue outbreak cases that departed significantly from seasonal norms. We then extended this model to perform true prediction by limiting the MCMC analysis

to only the information known up until the time of prediction, followed by an auto-regressive moving average (ARMA) model to extrapoate prediction values. All of these models included explicit representations of the distributions of variability in both the internal and output model variables. The combined MCMC/ARMA model showed effective prediction for data from San Juan, Puerto Rico, and somewhat less effective prediction for data from Iquitos, Peru. We also studied the problem using an ensemble of decision trees approach. This work examined the question of whether the provided contextual data inherently incorporated information that correlated with dengue fever outbreaks, and which contextual variables were most correlated.

The details of our dengue fever study are described in a conference paper submission [9]; see below.

2. *Wind power.*
We obtained a data set from Bonneville Power Administration (BPA) listing power production and associated forecasts for 33 wind power production facilities in the Columbia River Basin. These data list hourly wind power production, and also associated forecasts with various forecast lead times. The forecasts were provided to BPA by a commercial vendor. We first defined events of interest to be rapid increases and decreases of power (called "ramps"), but decided that developing a method to forecast ramps would be redundant with ramp information already present in the forecasts. We also observed that while forecasts typically captured the structure of wind ramps, they sometimes suffered a temporal offset, where the forecast was either early or late with respect to the true ramp. These discrepancies cause economic loss, and provided an opportunity for study: Can we predict these offsets? Can we use offset estimation techniques to improve economic performance?

   We approached this problem by developing new algorithms to estimate local time shifts that would improve overall forecast accuracy. As with the dengue fever study, our first analysis considered the entire data set. This full-time retrospective analysis indicated that selection of optimal time shifts could lead to significant improvement in economic performance throughout the year. This was essentially a model fit estimation problem. Then we extended this analysis to a true prediction model, where only information available at the time of prediction was considered. As with the dengue fever problem, this increased difficulty significantly. Using the true prediction model, the estimated economic benefit was reduced, and the effective lead time was greatly reduced. Note that both the dengue fever and wind power studies drove home the substantial increase in difficulty that results from modifying a full-time retrospective analysis to a predictive analysis restricted only to contemporaneously-availalble information.

   We also studied geospatial relationships between locations, to gain insight into whether or not graph relationships might improve predictive performance. Our study identified significant lead/lag correlation in the weather itself, which was also well-represented in the forecasts. But we did not observe significant non-local correlations in the *error* in the forecasts, suggsting that graph-based constraints would not yield signfiicant improvement in temporal shift estimation. The fact that geospatial correlations were significant in the raw weather but not forecast error is not surprising, since weather

is a physical phenomenon, and forecast error is a function of the relationship between the physical weather and human-designed algorithms to model the weather.

The details of our wind power study were reported in presentations at two workshops [12, 13] and a longer detailed report [14]; see below.

3. *Network analysis.*
We began with a network traffic data set from a real computer network with associated events, but ultimately determined that we did not understand the data's provenance sufficiently to resolve questions that arose. This led us to look for other data solutions. One promising example was the LBNL/ICSI Enterprise Tracing Project data set from Lawrence Berkeley National Laboratory (LBNL) [8]. This data set had realistic network data, but did not have correlated external events of interest. A second approch would be to use synthetic data carefully constructed to incorporate realistic network traffic and response to external events; this led to interactions with Madhav Marathe's group at Virginia Tech exploring possible data and analysis collaborations.

   Our analysis approach to this problem was to apply Ripley's K-Function [6]. This function was developed to study spatial patterns; we extended it to apply to temporal patterns, and further extended it to amplify the signal-to-noise ratio when seeking systematic precursors. While initial progress was encouraging, unfortunately we had to set aside this problem to favor complete studies of the other challenge problems, and ran out of time before we could get back to it.

   A description of our work in progress on the network analysis problem is given in Section 3.1.

For a detailed description of the dengue fever and wind power prediction results, see the publications and reports cited in the next chapter. For a detailed description of the network precursor analysis work in progress, see Section 3.1.

1.4 Lessons Learned

There were several key lessons learned through the course of our work:

- General-purpose prediction is a very difficult problem.
- It is essential to ensure that prediction calculations only make use of information available at the time of prediction. This can be subtle to implement. However, enforcing this limit can dramatically reduce calculation accuracy, due to the underconstrained nature of forward time extrapolation. Mathematically, "Hindsight is 20/20."
- Prediction without domain knowledge is exceedingly difficult. Successful prediction methods that exploit domain knowledge are well-known, such as classical physics, which routinely predicts the performance of physical systems. This is because the physical

system has inherent structure which constrains its behavior, and these constraints are captured in our models of these processes. It is noteworthy that even in this well-understood domain, prediction sometimes fails when aspects of the physical system fail to provide sufficient constraint; examples include predicting the head/tail outcome of a bouncing coin, and the detailed behavior of unconstrained fluid jets.

- Precursor identification in the presence of noise is confounded, because the precursor signal may be smaller in magnitude than the effects explainable by normal variability.

- Geospatial-temporal datasets are often inadequate for statistical prediction analysis. We found many initially promising data sets, only to learn that they lacked either spatial density, temporal consistency, or both. This can be explained by economic drivers: First, deploying a wide array of sensors and operating them continuously for a long time is expensive, which drives choices of either reduced spatial sampling, or repositioning sensors to obtain more sample positions at the expense of temporal continuity. Second, full data sets containing both high spatial density and complete temporal sampling are large, leading to storage and maintenance expenses. Without a clear need or method to process such data, it is difficult to justify the expense of collection and curation.

- To improve prediction performance, obtain datasets that are simultaneously spatially and temporally dense, and apply system models and domain knowledge to improve predictive accuracy. Note that a well-known example where both of these are applied is weather prediction – a high-value enterprise with both spatially and temporally dense data archives, whose analysis is supported by rich and detailed physical models.

1.5 Path Forward

We envision several potential opportunities for future work:

1. *Dengue fever.* Extending the analysis to consider other weather variables, such as pressure, might yield some improvement. Pressure is mentioned here because the decision-tree analysis suggested that it is correlated with dengue cases. Beyond this, it could be fruitful to apply the technique to other regions with dengue fever, or even to other diseases, to gain an understanding of the robustness of the approach across a range of examples. It would be also of great interest to study the method's applicability to other types of prediction problems.

2. *Wind power:* The shift estimation algorithms we developed in our wind study only considered temporal shift values that were integer hours; it might improve performance to extend this technique to consider fractional hour shifts. This might improve economic benefit and might also affect the significance of geospatial relationships. However, it's not clear this is worthwile given the lack of a domain model of human forecast error,

and the impermanance of forecast error characteristics resulting from revisions to the forecast algorithms over time.

3. *Network analysis:* The K-function analysis technique shows intriguing potential for a variety of problems. A key next step would be to explore how the precursor identification method performs in the presence of variable lead time, where the lead time variability is characetrized by either (a) simple noise in similar lead times, or (b) by qualitiatively different lead times, which might in turn be related to schedule or trigger events. An additional next step would be to test the method using realistic network traffic data, obtained from wither LBNL or Virginia Tech as described above. Further valuable studies include characterizing the number of event repetitions required for successful K-function identification, possibly as a function of ambient noise levels.

4. *General:* As noted in Section 1.4, prediction using generic statistical approaches is much more difficult than prediction informed by domain knowledge. It could be very valuable to explore how to incorporate domain knowledge in time series analysis to produce model-informed analysis of statistical properties.

Chapter 2

Outcomes

In this section we list the publications, reports, and presentations that were produced in this project. These provide detailed descriptions of the project's primary results.

- A. Staid, Predicting Wind Power Ramp Events. Presented at INFORMS 2016 [12].

  Abstract:
  Wind power ramp events (large changes in output over a short period of time) are of particular concern in power systems with high wind penetration. They are also often difficult to predict. We present statistical methods for combining multi-source data to better predict the adverse ramp events that are typically not captured in a standard weather forecast. We present a case study using data from the Bonneville Power Administration and focus on farm-specific ramps.

- A. Staid and R. C. Brost. Data-Driven Approaches for Wind Power Ramp Timing at BPA. Presented at the *Utility Variable-Generation Integration Group (UVIG) Fall Technical Workshop,* October 2017 [13]. Presentation slides available on-line at https://www.uvig.org/.

- R. Brost, V. Leung, H. Link, C. Phillips, and A. Staid. Event Prediction Using Graph-Augmented Temporal Analysis. Poster in *2018 Conference on Data Analysis (CoDA),* March 2018 [2].

  Abstract:
  Temporal data streams can produce large volumes of data that are difficult to fathom. The observations issued from a single sensor over time contain temporal patterns that may reveal underlying periodic structure, or trigger-response relationships. These can often be superimposed, resulting in a cacophony of signals containing embedded information of interest. This challenge is compounded when multiple data streams have inter-stream relationships of importance. Our work seeks to analyze such data streams to find precursors that systematically precede events of interest. Because we anticipate that a purely temporal analysis might reveal numerous spurious precursors, we envision augmenting the temporal analysis with information drawn from a graph, which provides other relationship information to reinforce or deprecate relationships suggested by the pure temporal analysis. This work was inspired by previous results obtained using geospatial-temporal graphs applied to a variety of problems [3, 4]. Exploring this hypothesis, we address three example problems: prediction of Dengue fever outbreaks,

advance identification of errors in forecasts for wind power generation, and analysis of synthetic network traffic preceding events of interest. We explore three different temporal analysis approaches for these domains: Bayesian Markov-Chain Monte Carlo analysis [7], time shift analysis [13], and an extended version of Ripleys K-function [6]. We will present our work in progress, including preliminary results and lessons learned to date.

- H. E. Link, S. N. Richter, V. J. Leung, R. C. Brost, C. A. Phillips, and A. Staid. Statistical Models of Dengue Fever. To appear in *AusDM 2018, The 16th Australasian Data Mining Conference,* November 2018 [9].

  Abstract:
  We use Bayesian data analysis to predict dengue fever outbreaks and quantify the link between outbreaks and meteorological precursors tied to the breeding conditions of vector mosquitos. We use Hamiltonian Monte Carlo sampling to estimate a seasonal Gaussian process modeling infection rate, and aperiodic basis coefficients for the rate of an "outbreak level" of infection beyond seasonal trends across two separate regions. We use this outbreak level to estimate an autoregressive moving average (ARMA) model from which we extrapolate a forecast. We show that the resulting model has useful forecasting power in the 6-8 week range. The forecasts are not significantly more accurate with the inclusion of meteorological covariates than with infection trends alone.

- A. Staid and R. Brost. Data-Driven Wind Power Forecast Improvement Using Temporal Offsets. Working paper for journal submission [14].

  Abstract:
  Wind power forecasts are used in a wide range of power system planning decisions, and accurate forecasts allow for more efficient operations, increased revenue, and an increased utilization of wind power for electricity generation. Continuous improvements in forecast accuracy has been ongoing, with the majority of the methods used depending on advanced physics-based or statistical models. It can be observed in wind production data that shifts in the timing of large-scale events are often the cause of errors, as opposed to under- or over-predicting the amount of anticipated generation. Here, we present a purely data-driven method that aims to correct for temporal offsets in the timing of wind power ramp events by detecting leading indicators of mismatches in very short lead time settings. We apply this algorithm to a large dataset of 33 wind projects in the Bonneville Power Administration balancing authority. The results are promising for a subset of the wind projects, showing error reductions of 3-5% on average. This is the first attempt that we know of that focuses solely on temporal shifts in wind power forecasts.

Chapter 3

Work In Progress

This section describes work that has been performed under this project, but is not yet ready for publication. It documents what was accomplished, so that future work may build upon it. This work-in-progress is not reported in a stand-alone SAND report, because that might lead to confusion with potential later SAND reports written after additional work is performed to achieve a publishable state.

3.1 Discrete Event Precursor Analysis

Unlike problems which involve continuously-measured variables such as dengue fever rates or wind power production, network traffic analysis is characterized by discrete events. The question is not "How much traffic is there?" but rather "Are there specific traffic events which are precursors to events of interest?" As a result, a different analysis approach is required. In this work, we applied an extended version Ripley's K-function [6] to find precursors to events of interest in a simulated network traffic problem with synthetic data. While this represents significant progress, we feel that additional work is required before this reaches a state of completion suitable for publication. Thus in this section we describe our work in progress, for archival and reporting purposes.

Network traffic is characterized by discrete events – an event occurs each time a network message is sent. This is similar to the locations of discrete objects in space – a location is determined at each place that a discrete object is found. Ripley's K-function was designed to analyze spatial distributions of objects, in order to identify statistical properties that reveal structure to their placement. This is well described by Dixon in [6]. Dixon presents an example of analyzing the locations of trees, and comparing their placement against a hypothesis of random noise. In the sections that follow, we will first describe the properties of Ripley's K-function on spatial examples, and then describe our extensions to temporal problems and an example network analysis problem.

3.1.1 Ripley's K-function

Ripley's K-function is a function of radius r, and measures the expected value of the number of objects within a circle of radius r placed anywhere in the domain. This is illustrated in Figure 3.1. Conceptually, imagine placing a circle of radius r somewhere in the field of

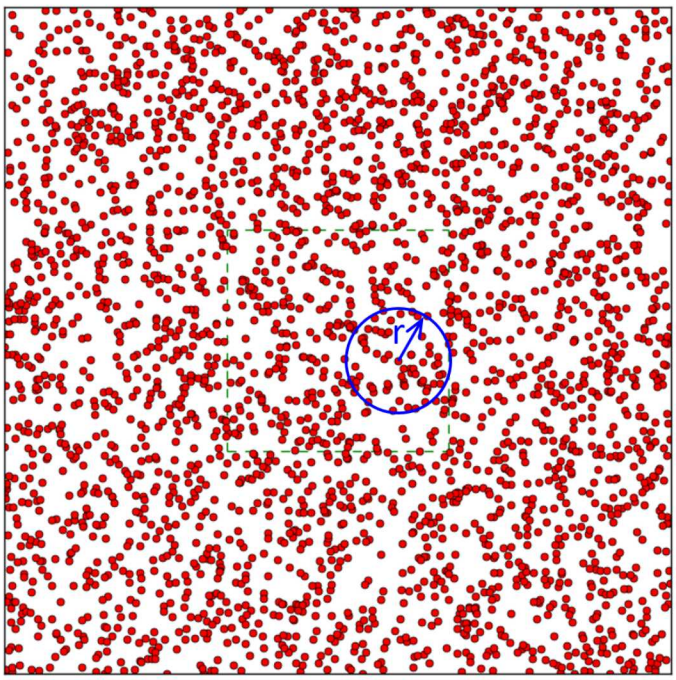

Figure 3.1. For a given radius r, how many points are expected on average?

interest. Then count the number of points within the circle. Repeat this experiment for many different circle positions, and compute the mean of the resulting counts. This is the K-function value for this particular r. To establish the K-function over a range of r values, simply repeat this process for each r value of interest.

This conceptual experiment actually measures the value of $\hat{K}(r)$, an estimate of the true underlying $K(r)$ function. For the remainder of this discussion, we will describe techniques for computing the estimate $\hat{K}(r)$; analysis of $K(r)$ is beyond the scope of this report.

Given a set of N sample points in some spatial domain, we can compute $\hat{K}(r)$ using the following equations:

$$\hat{\lambda} = \frac{N}{A} \tag{3.1}$$

$$\hat{K}(r) = \hat{\lambda}^{-1} \sum_{i=1}^{N} \sum_{j \neq i}^{N} \frac{I[d_{ij} < r]}{N} \tag{3.2}$$

where N is the number of points, A is the area of the region of interest, d_{ij} is the distance between point i and point j, and I[predicate] is the indicator function that returns 1 if the predicate is true, 0 otherwise.

Note that Dixon addresses the issue of edge effects in [6]. Returning to the previous conceptual experiment, imagine circle positions where a portion of the circle falls outside the square boundary. Obviously the number of points in the circle will be reduced. If this is not considered, then an abundance of such placements will distort the overall computation, resulting in an underestimate of $\hat{K}(r)$. Dixon presents a correction factor to solve this problem, resulting in an equation that is more complex than Equation 3.2. We avoid this complexity by instead modifying the domain to eliminate edge effects. In our procedure we restrict circle placements to an inner region that has a surrounding margin wide enough to ensure that no crcle placement has any portion that falls outside the overall region boundary. This restricted region is shown in Figure 3.1 as a dashed green inner square. This simplifies the analysis both numerically and conceptually, and is valid as long as r does not exceed the margin width.

Continuing to follow Dixon, we then construct a transformed function estimate $\hat{L}(r)$, tailored to a comparison hypothesis of complete spatial randomness:

$$\hat{L}(r) = \sqrt{\frac{\hat{K}(r)}{\pi}} \tag{3.3}$$

This normalizes $\hat{K}(r)$ so that, if the actual point distribution is complete spatial randomness, then in the limit $\hat{L}(r) \to r$, for all r. This allows us to examine the following comparison function to assess whether the given points follow complete spatial randomness:

$$\hat{L}(r) - r \tag{3.4}$$

Figure 3.2 shows plots of $\hat{K}(r)$, $\hat{L}(r)$, and $\hat{L}(r) - r$ for an ideal set of many points placed with perfect spatial randomness. The key result is the bottom plot, where $\hat{L}(r) - r$ follows the horizontal axis. Deviations of $\hat{L}(r) - r$ from the axis indicate deviations of the point distribution from perfect spatial randomness.

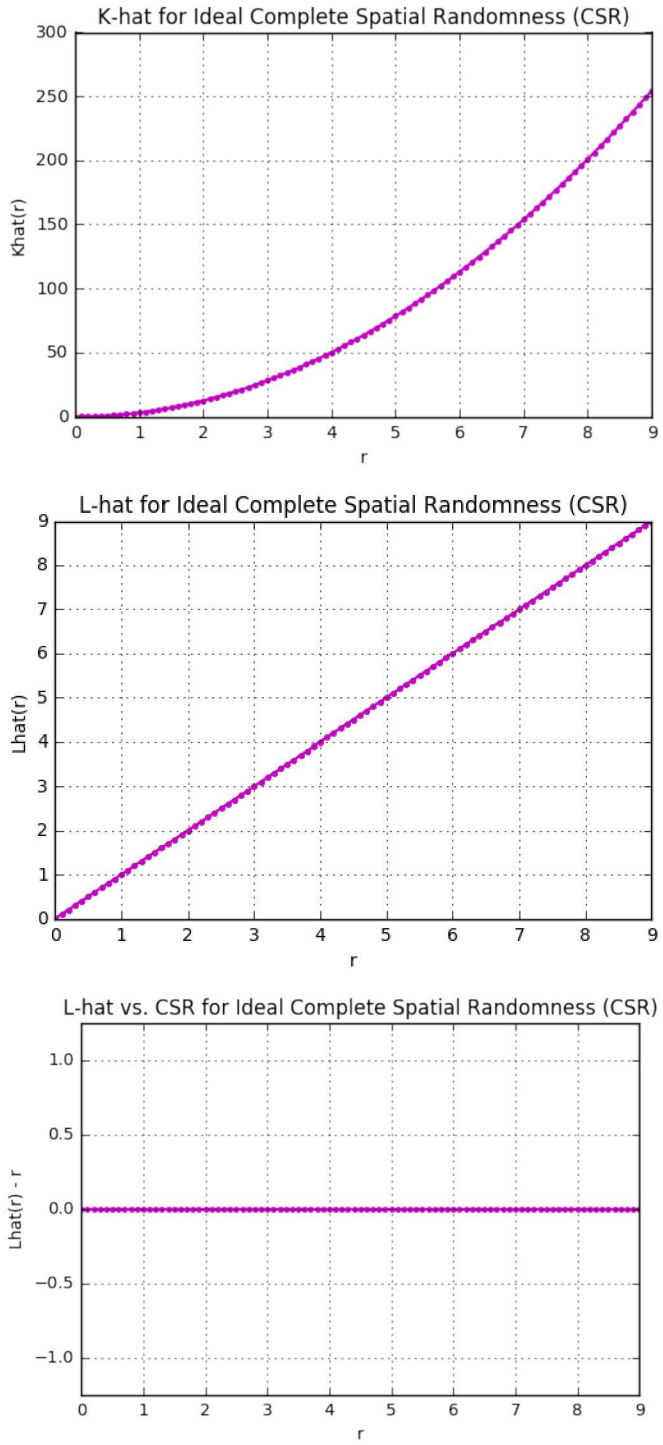

Figure 3.2. Plot for cases with complete spatial randomness. Top: $\hat{K}(r)$. Middle: $\hat{L}(r)$. Bottom: $\hat{L}(r) - r$.

The value of this function for revealing structure in the input point set is illustrated in Figure 3.3.

The top of Figure 3.3 shows a randomly generated set of points, and the corresponding plot of $\hat{L}(r) - r$. The curve is close to the horizontal axis, but does not follow it perfectly, due to the limited number of sample points, and the imperfection of this particular sample compared to complete spatial randomness.

The middle of Figure 3.3 shows a regular grid of points, and the corresponding plot of $\hat{L}(r) - r$. Note the significant structure in the resulting curve, and its substantial deviation from the horizontal axis. The domain $r \in [0, 1)$ is especially prominent, because on a regular grid the number of points of distance less than one is zero, for all points. The jagged teeth of this curve can be attributed to the critical distances where additional points suddenly come within range, as illustrated by the accompanying diagrams.

The bottom of Figure 3.3 shows a comparison between the random and grid points. The thin black curves correspond to a different sample of random points; each curve is analogous to the red curve in the top of the figure, but for a different random sample. There are 55 curves shown, corresponding to 55 different random point samples. We can view the shape spanned by the 55 thin black curves as an estimate of the range of values that might be attained by $\hat{L}(r) - r$ for a given point sample. In other words, values of $\hat{L}(r) - r$ that fall within this envelope can be readily explained as instances of ordinary random variation. Note that the width of the envelope grows as r increases.

The blue curve shown in the bottom of Figure 3.3 is the $\hat{L}(r) - r$ curve from the grid example in the middle of the figure. Note that the strong jagged peaks at $r = 1$, $r = \sqrt{2}$, and $r = 2$ are well outside the envelope, indicating that the grid sample point locations have structure that departs substantially from random noise. This makes sense, given that the regular grid is not random. Further, for larger values of r the curve lies inside the variation envelope, indicating that for these distances, it is difficult to distinguish the grid from random point placement. We can get some intuition of this by imagining viewing the two cases from a far distance. From a distant view point, it would be much more difficult to distinguish random point placements from a regular grid.

Thus Ripley's $\hat{K}(r)$ function provides a means of identifying structure in a sample of locations. Dixon presents an interesting example of this in [6]. His analysis indicates a dearth of trees at very close distances (due to trunk interference and root competition), and also an excess of trees at moderate distances (due to clustering effects tied to the tree's reproductive mechanism). This raised our interest in applying Ripley's K-function to temporal problems, hoping to deduce structure within a sampled time series of discrete events.

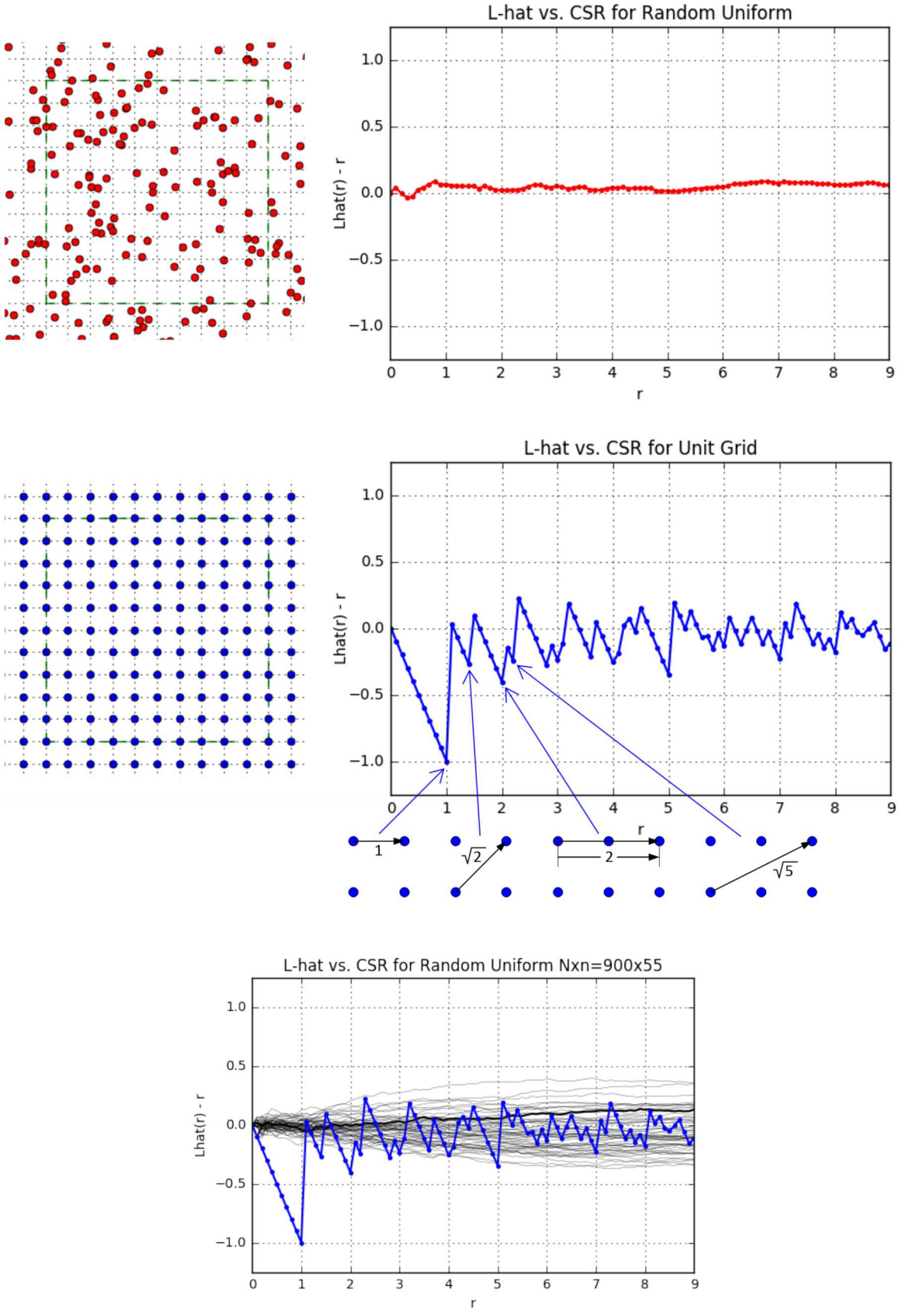

Figure 3.3. Random points vs. a regular grid.

3.1.2 Extension to Temporal Analysis

We constructed an extended version of Ripley's K-function for temporal analysis of discrete events, comparing a given background event stream B against a list of events of interest V:

$$\hat{\lambda} = \frac{N_B}{T} \tag{3.5}$$

$$\hat{K}_{2\text{pre}}(dt) = \hat{\lambda}^{-1} \sum_{v}^{N_V} \sum_{b:t_b<t_v}^{N_B} \frac{I[(t_v - t_b) < dt]}{N_{B:t_b<t_v}} \tag{3.6}$$

$$\hat{L}(dt) = \hat{K}_{2\text{pre}}(dt) \tag{3.7}$$

where N_B is the number of events in the background data stream B, T is the total time span, N_V is the number of events of interest V, and t_v and t_b are event times. Equation 3.6 ensures that the term $(t_v - t_b)$ is always positive, by considering only background event times t_b that are earlier than the current event time of interest t_v.

Since time is a linear domain rather than a two-dimensional domain, the computation of $\hat{L}(dt)$ has a different structure than $\hat{L}(r)$. The calculation of $\hat{L}(dt)$ shown in Equation 3.7 is tailored to complete temporal randomness. Given this, the comparison function we plot has the same structure as the spatial case:

$$\hat{L}(dt) - dt \tag{3.8}$$

If the temporal distribution of the given event sample matches complete temporal randomness, then the plot of $\hat{L}(dt) - dt$ should lie on the horizontal axis. Deviations indicate either a structural deviation from temporal randomness, or an imperfection in the sample.

3.1.3 Temporal Examples

Figure 3.4 shows three example discrete event streams. In all of these, time increases from left to right, and each vertical bar corresponds to a different event. The top example shows an event sequence with a constant time gap Δt between events. The middle case is an example of complete temporal noise; event times are uniformly distributed. In the bottom case, a regular sequence of events occurs with time gap Δt, and each event is preceded by a precursor with a lead time of dt_{pre}.

Figure 3.5 shows the results of applying Equation 3.6, where the top event stream in Figure 3.4 contains the events of interest. The top plot of Figure 3.5 shows the events of

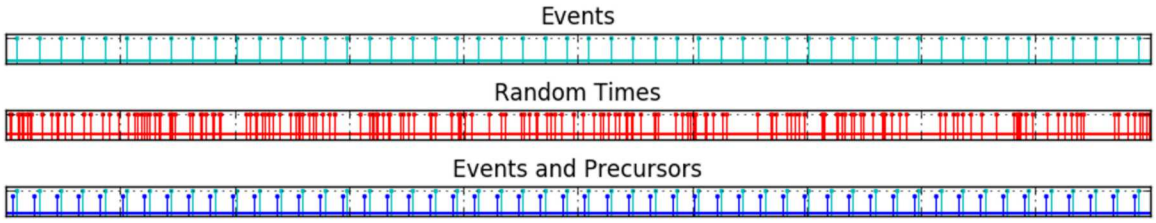

Figure 3.4. Example event traces.

interest compared to the temporally random events in the middle of Figure 3.4, while the bottom plot shows the events of interest compared to the dark blue precursor events in the bottom trace of Figure 3.4. The curve in the top plot roughly follows the horizontal axis, indicating temporal random noise as expected. The curve in the bottom plot shows jags that result from the structure of the precursors which precede each event of interest by exactly dt_{pre}, and the regular spacing Δt of the events. The jagged structure of the bottom plot is reminiscent of the regular grid plot shown in Figure 3.3, but is simpler since the time sequence is one-dimensional.

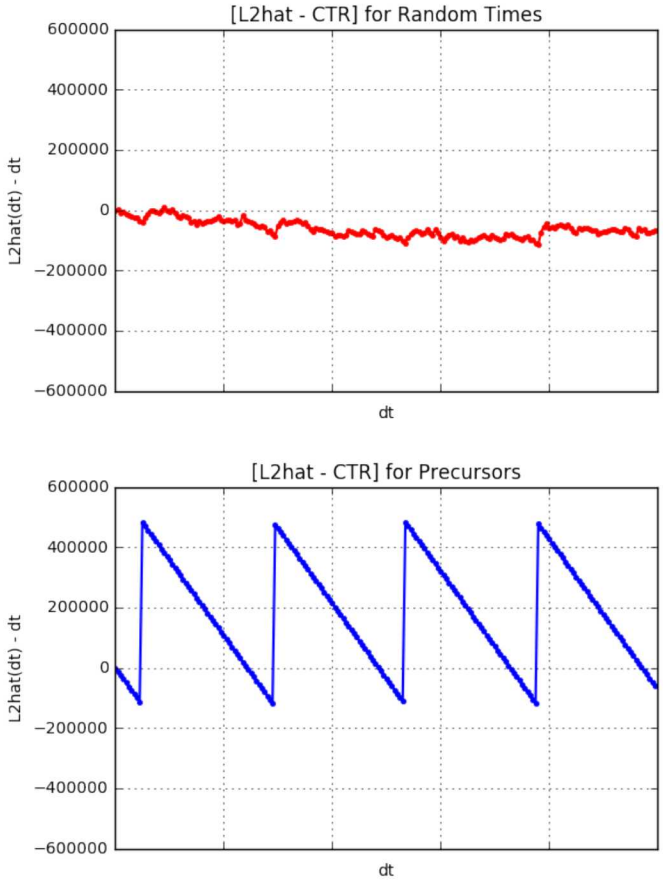

Figure 3.5. Plots of $\hat{L}(t) - t$. Top: For an example random event trace. Bottom: For regular events with precursors.

Figure 3.6 shows the effect of systematic precursors embedded within a random event stream. The top of the figure illustrates a regular sequence of events of interest compared against an random event sequence, with no relationship between the events of interest and the random event stream. The $\hat{L}(dt) - dt$ curves in the upper plot look like ordinary random noise. The thin black lines represent repeated samples of the random event stream; these delineate an envelope of variability similar to the one seen in Figure 3.3. Instances of $\hat{L}(dt) - dt$ that fall within this envelope are not distinguishable from random noise.

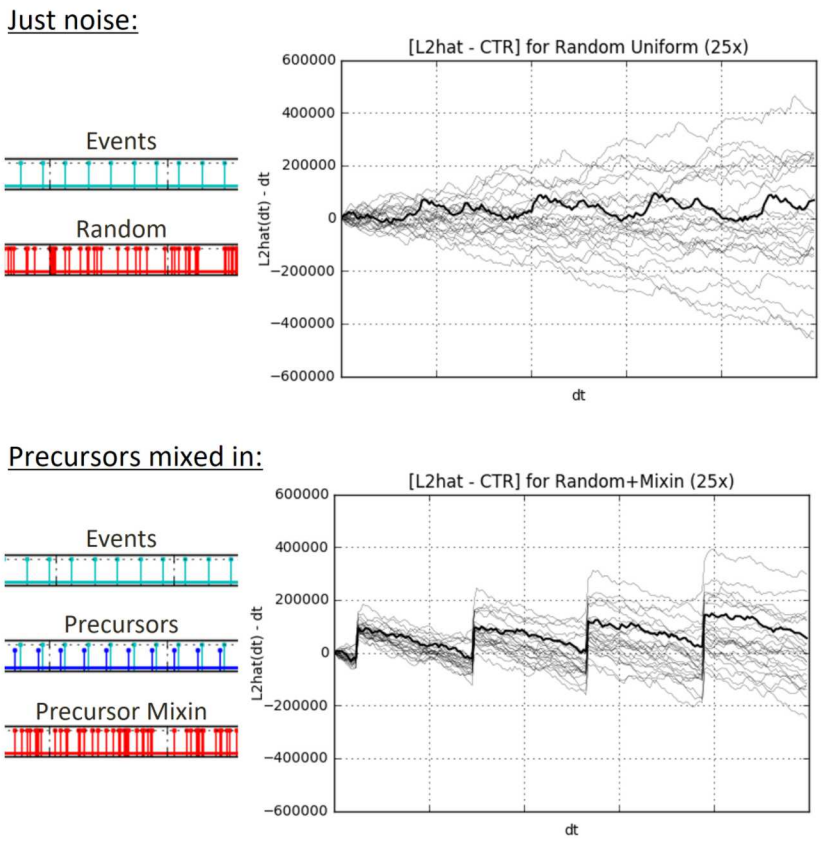

Figure 3.6. Comparison of analysis of events against random background vs. random background with precursors mixed in.

The bottom of Figure 3.6 illustrates the same random event stream, but with precursors to the events of interest added, with a fixed precursor lead time of dt_{pre}. The $\hat{L}(dt) - dt$ curves now exhibit a combination of jagged and random structure. Analysis of the various cases reveals that the sharp vertical rises are due to the precursors, and are spaced according to the precursor lead time dt_{pre} and the event of interest spacing Δt.

Can the precursor be identified apart from the random noise? Overlaying the envelopes with and without precursors reveals that for this example, the first dt_{pre} cycle with precursors is likely to lie outside the envelope of variability of pure temporal randomness. But the second cycle is only marginally distinguishable, and the third and further cycles lie well within the noise variation regime. And this is for a case with short, perfectly repeatable precursor lead time. We anticipate that as precursor lead time increases and incorporates noise, $\hat{L}(dt) - dt$ will become increasingly difficult to reliably distinguish from random variation.

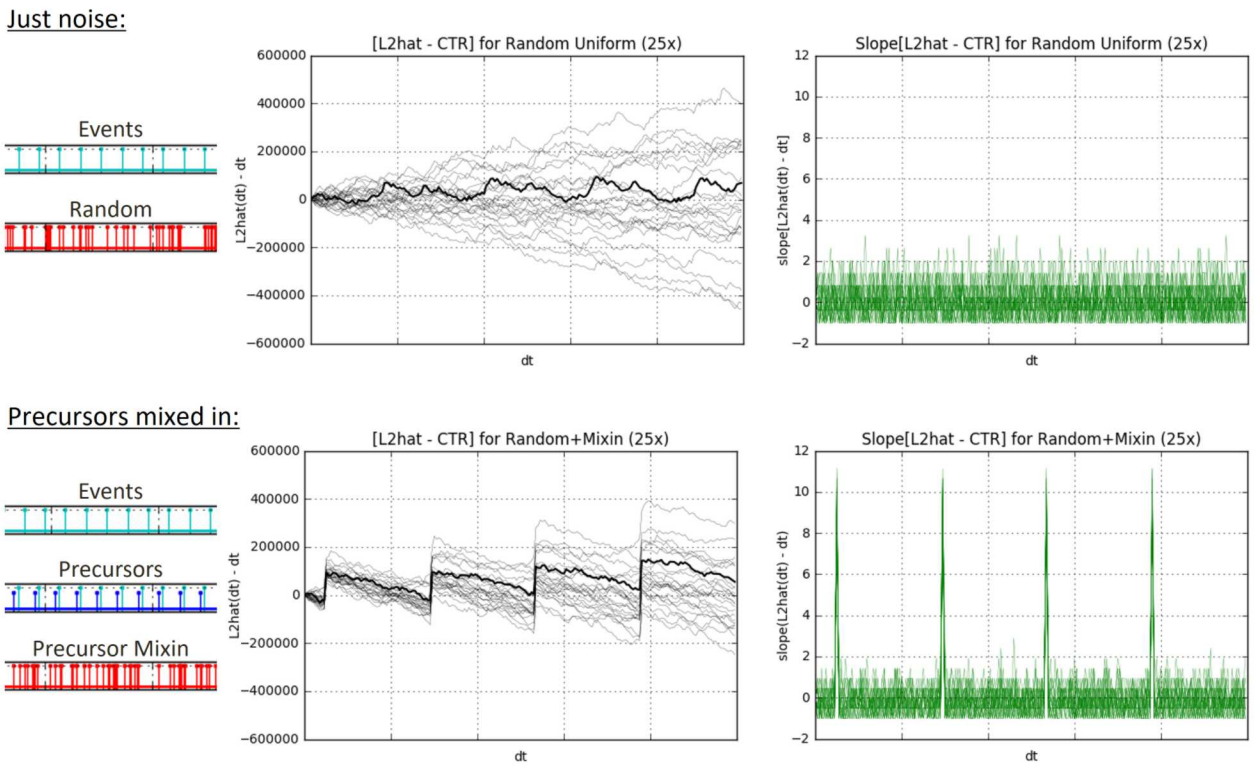

Figure 3.7. Comparison of Figure 3.6, showing result of slope analysis.

3.1.4 Amplifying Signal-to-Noise

Yet the jagged structure of the $\hat{L}(dt) - dt$ curve seen in the lower plot Figure 3.6 clearly contains structure corresponding to the presence of event precursors. Can we exploit this structure? One approach would be to compute the slope of the $\hat{L}(dt) - dt$ curve, and compare that against noise instead of simply checking the magnitude of $\hat{L}(dt) - dt$.

Figure 3.7 shows the result of this approach. In the top right, we plot the slope of $\hat{L}(dt) - dt$ for the case of pure temporal randomness. The result is a uniform noise level. In the bottom right, we see how the slope of $\hat{L}(dt) - dt$ is influenced by the presence of precursors. Strong peaks are seen, the first at the time dt_{pre}, and subsequently appearing at intervals of $dt_{\text{pre}} + \Delta t$. These peaks are easily distinguished from the background noise level.

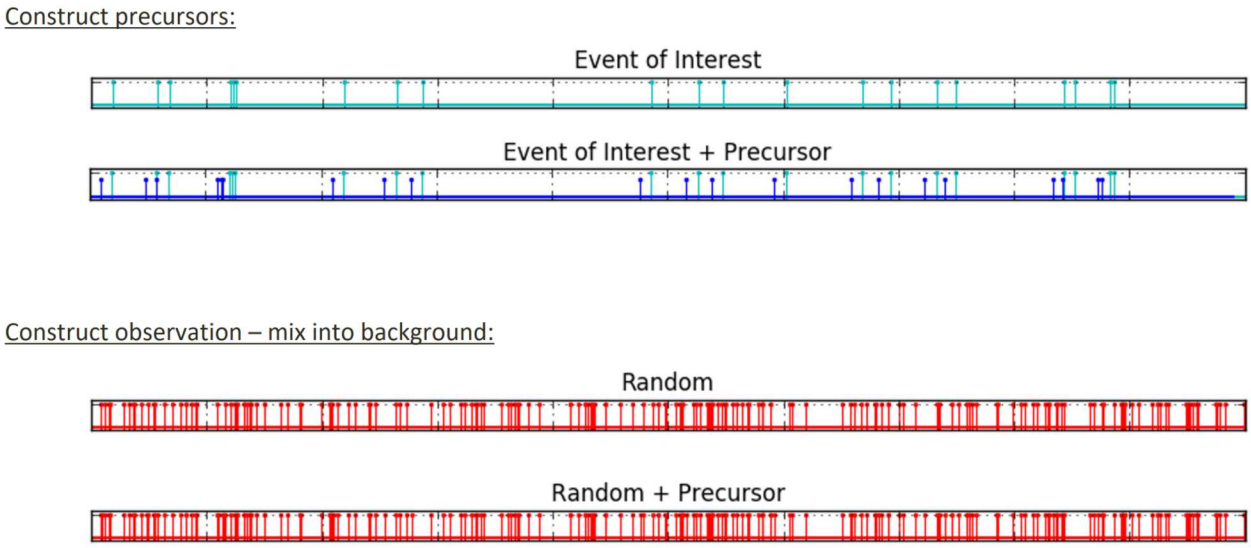

Figure 3.8. Setting up a precursor analysis example.

Figures 3.8 and 3.9 show this approach applied to a case where the events of interest are not evenly spaced, and where the precursor lead time is still a constant dt_{pre}.

Figure 3.8 shows the problem setup. First a sequence of events of interest is defined, with uneven spacing. Next precursors are constructed with a constant precursor lead time dt_{pre}. Next a random background event stream is generated. Finally, the precursors are added to the random event stream. To the naked eye, they are completely hidden within the noise.

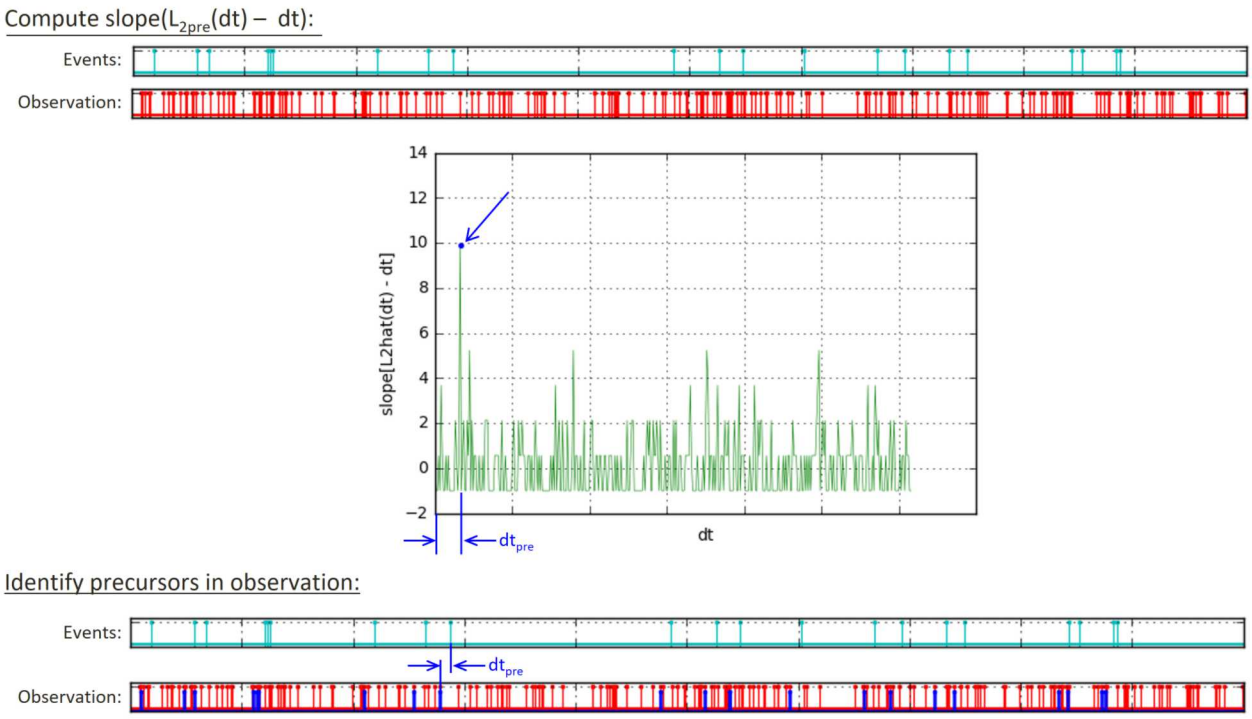

Figure 3.9. Precursor analysis with signal amplification.

Figure 3.9 shows our analysis of this problem. The slope of $\hat{L}(dt) - dt$ is plotted with a clear peak exhibited at dt_{pre}. This provides the detection signal indicating that precursors are present, and their specific lead time. We can then use this lead time to find the specific precursor events embedded within the noisy data stream. By visiting each event of interest and looking backward a time interval of dt_{pre}, we can identify each precursor. These are shown in blue at the bottom of Figure 3.9. Finding specific precursor events enables further investigation of attributes of those events, or search for contemporaneous related information.

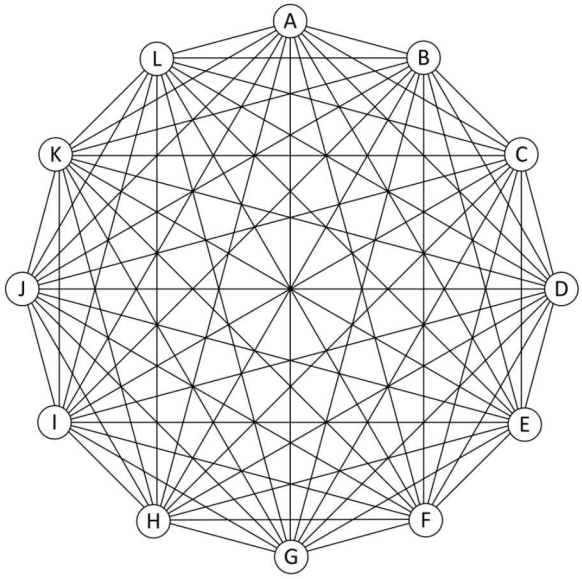

Figure 3.10. A hypothetical network with 12 nodes.

Figure 3.11. History of pizza deliveries.

3.1.5 A Simulated Network Traffic Example

In this section we will present a hypothetical example demonstrating the use of the extended K-function. We consider the following network traffic analysis problem:

> There is a small company with 12 employees. On seemingly random occasions, a group of employees gathers in the break room for lunch, with several piping hot pizzas delivered. Our hero, a sysadmin in the group, sees these lunches and would like to join in. But they don't know whom to ask for an invitation. So instead of simply asking around, our socially shy sysadmin decides to solve the problem by examining the network traffic. Who orders the pizza?

Our sysadmin solves this problem by considering the local office network as a graph, shown in Figure3.10. Each graph node corresponds to a different employee's computer, and each edge corresponds to computer-to-computer communication links. A message exchanged from one computer to another is an event associated with that edge.

Our sysadmin also knows the history of when pizza lunches happened in the past. This log of events of interest is shown in Figure 3.11.

Figure 3.12. Pizza deliveries vs. all traffic.

Figure 3.12 shows the log of network traffic for all of the machine-to-machine edges. For this hypothetical example, we simply generated random noise for the traffic on each edge. Is there a pattern that can be found amidst this cacophony of random traffic?

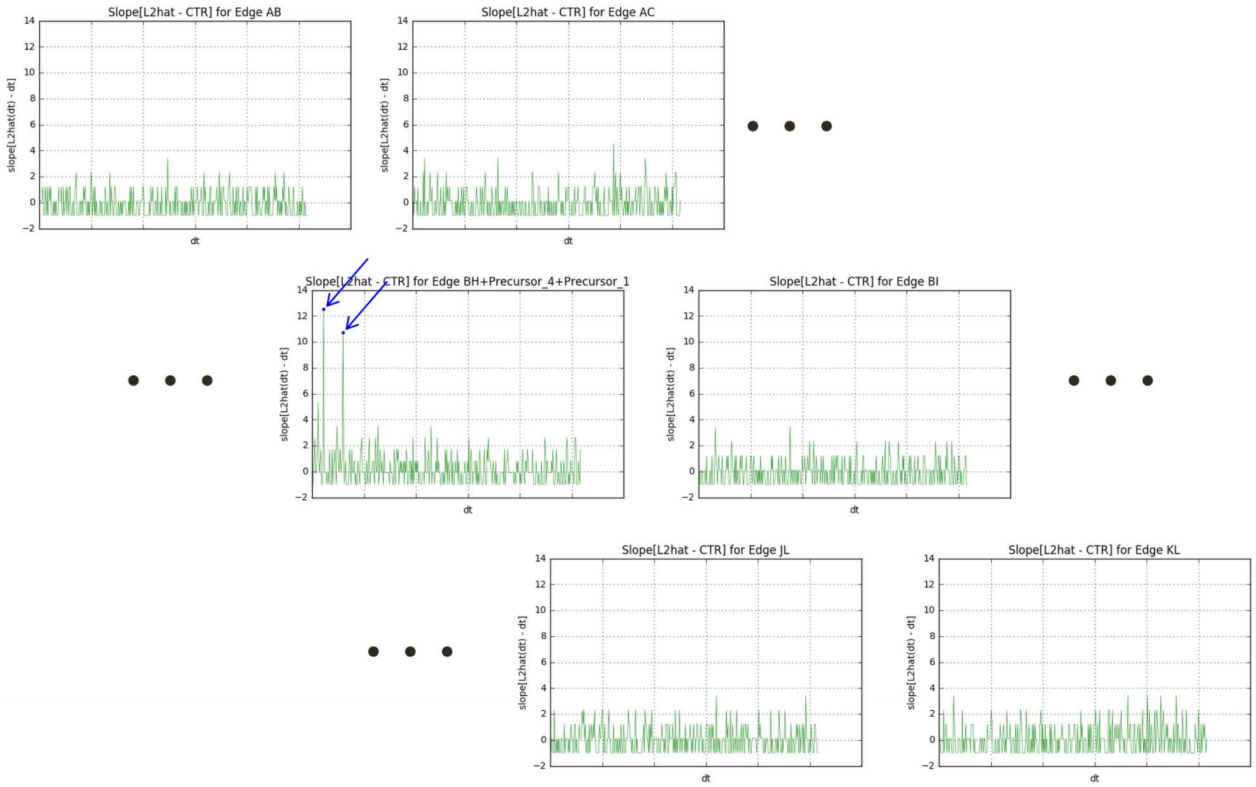

Figure 3.13. Temporal traffic analysis.

Figure 3.13 shows the analysis of this problem. For each network edge, a temporal analysis is performed following the technique shown in Figure 3.9. For most edges only a low-level noise signal is found. But for a few edges, a strong precursor signal is identified. The blue arrows in Figure 3.13 indicate an example.

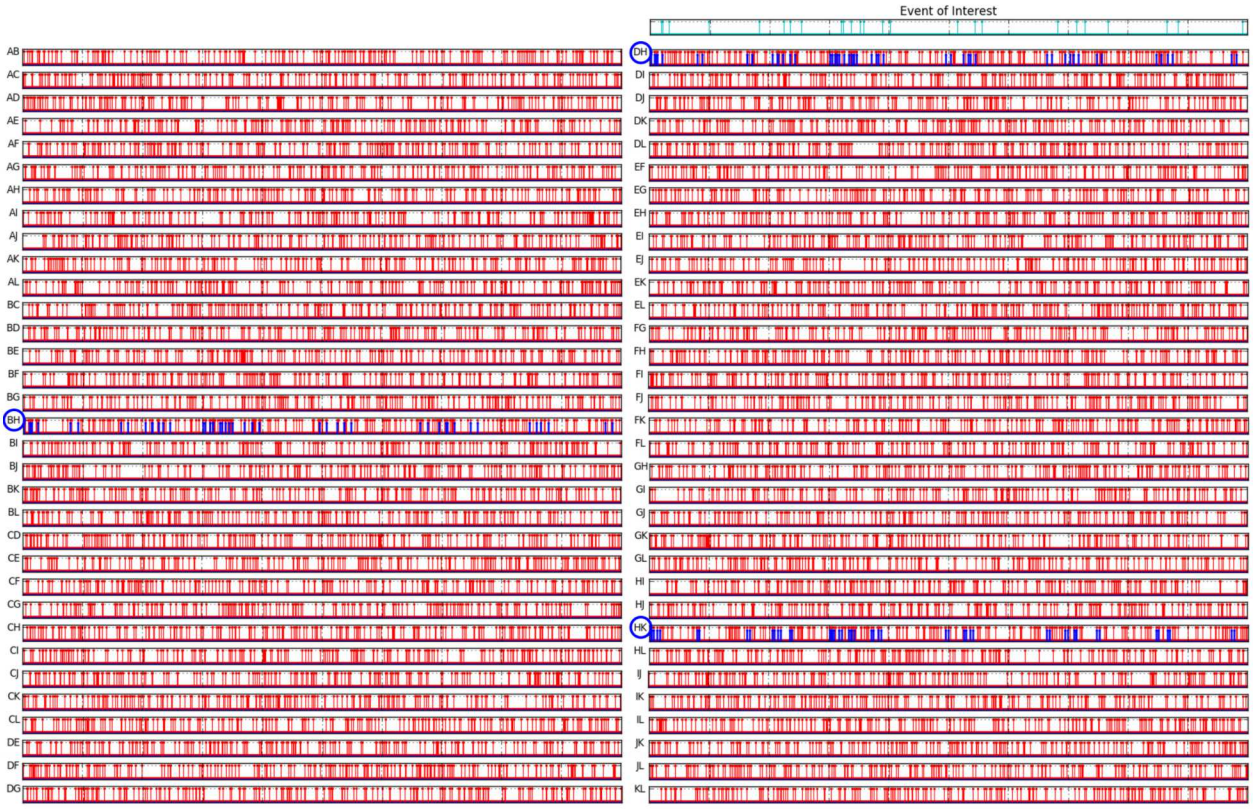

Figure 3.14. Precursors to pizza deliveries found.

Figure 3.14 shows the precursors found by the method, rendered in the context of the original network traffic. We can see that for three edges — BH, DH, and HK — precursors reliably appear just prior to every pizza delivery. These edges, and the computers that define them, are clearly associated with the pizza lunches.

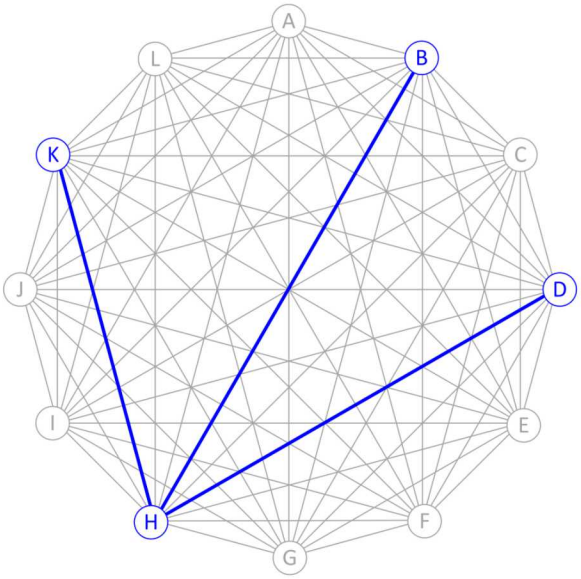

Figure 3.15. Precursor analysis results, rendered in the network graph.

Figure 3.15 shows these results rendered in the context of the network graph. We see that node H is common to all of the communications. What's happening is that nodes K, B, and D communicate pizza topping requests to node H, who then places the order. Our sysadmin concludes that the owner of node H is the person to ask about the pizza lunches. Now if they could only get up enough nerve...

This is an example of graph-augmented temporal analysis. By combining temporal analysis with graph constraints, we were able to identify a solution that could not have been found by either method working alone. Despite the overwhelming confusion in the overall network traffic data stream, the extended Ripley's K-function was able to identify systematic precursors, and viewing the results in the graph clarified the key participating node.

However, this method is known to be brittle. Our example includes precise, consistent precursor lead times. Nodes K, B, and D communicate their topping orders to node H with exactly the same lead time before each pizza delivery. In a more realistic example these lead times would exhibit variation, and this variation would in turn weaken the precursor signal in the temporal analysis. Understanding the degree of this weakening and developing countermeasures to preserve the signal in the face of lead time variation remain topics for future work. This is further discussed in Section 1.5.

References

[1] Brandon Behlendorf, Gary LaFree, and Richard Legault. Microcycles of violence: Evidence from terrorist attacks by eta and the fmln. *Journal of quantitative criminology*, 28(1):49–75, 2012.

[2] Randy C. Brost, Vitus J. Leung, Hamilton E. Link, Cynthia A. Phillips, and Andrea Staid. Event prediction using graph-augmented temporal analysis. Poster at 2018 Conference on Data Analysis (CoDA), March 2018.

[3] Randy C. Brost, William C. McLendon-III, Ojas Parekh, Mark D. Rintoul, David R. Strip, and Diane M. Woodbridge. A computational framework for ontologically storing and analyzing very large overhead image sets. In *3rd ACM SIGSPATIAL International Workshop on Analytics for Big Geospatial Data (BigSpatial-2014)*, November 2014.

[4] Randy C. Brost, David N. Perkins, and Kristina Czuchlewski. Activity analysis with geospatial-temporal semantic graphs. Sandia Report SAND2015-10455, Sandia National Laboratories, December 2015.

[5] Zachary D Danks and William F Porter. Temporal, spatial, and landscape habitat characteristics of moosevehicle collisions in western maine. *The Journal of Wildlife Management*, 74(6):1229–1241, 2010.

[6] Phillip M. Dixon. Ripley's k function. *Encyclopedia of Environmetrics*, 3:1796–1803, 2002.

[7] John K. Kruschke. *Doing Bayesian Data Analysis: A Tutorial with R, JAGS, and Stan, 2nd Edition.* Elsevier, 2015.

[8] Lawrence Berkeley National Laboratory. Lbnl/icsi enterprise tracing project. https://www.icir.org/enterprise-tracing/Overview.html.

[9] Hamilton E. Link, Samuel N. Richter, Vitus J. Leung, Randy C. Brost, Cynthia A. Phillips, and Andrea Staid. Statistical models of dengue fever. In *AusDM 2018, The 16th Australasian Data Mining Conference*, November 2018. To appear.

[10] NOAA. Combating dengue with infectious disease forecasting. Technical report, National Oceanic and Atmospheric Administration, DOC, June 5 2015. Retrieved from Dengue Forecasting http://dengueforecasting.noaa.gov/.

[11] The GDELT Project. The gdelt project. https://www.gdeltproject.org/.

[12] Andrea Staid. Predicting wind power ramp events. In *INFORMS 2016*, November 2016.

[13] Andrea Staid and Randy C. Brost. Data-driven approaches for wind power ramp timing at bpa. In *Utility Variable-Generation Integration Group (UVIG) Fall Technical Workshop*, October 2017.

[14] Andrea Staid and Randy C. Brost. Data-driven wind power forecast improvement using temporal offsets. Working paper for journal submission.

DISTRIBUTION:

1	MS 0899	Technical Library, 9536 (electronic copy)
1	MS 0359	D. Chavez, LDRD Office, 1171

Sandia National Laboratories