HIGH PERFORMANCE DATA ANALYTICS

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by **Battelle** *Since 1965*

Event Analysis and Recurrent Pattern Discovery

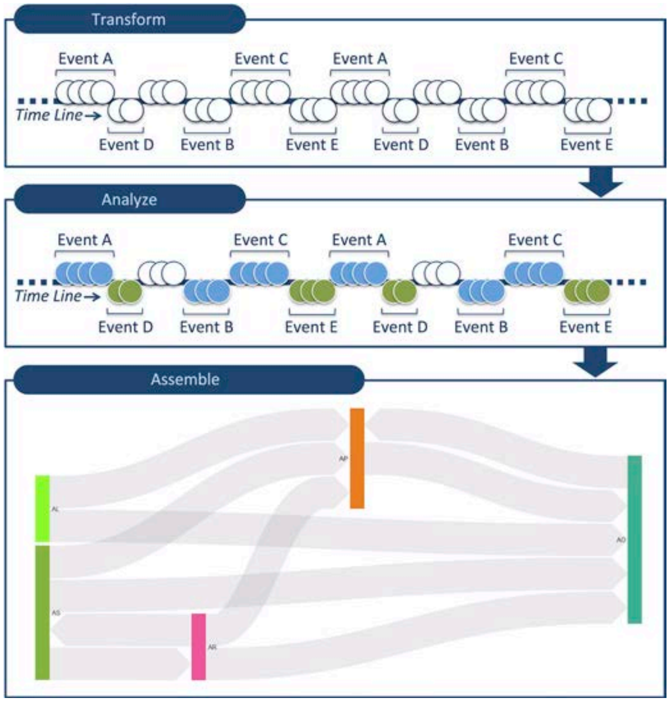
CHALLENGE

Causality describes how systems, processes, and components interrelate with one another. It is essential knowledge for sensemaking and decision-making–both elements of cyber situation awareness. Discovering causal structure from observations is daunting in cyberspace, a complex adaptive system comprising indefinite mixtures of overlapping processes. Processes are observed as sequences of events. System observers can discern events, not the true system state or the active processes. Co-occurring processes obscure causal relationships: at any given time, multiple processes may concurrently execute, including numerous instances of the same process. The objective is to decompose a mixture into its constituents by identifying causal structure within event sequences.

CURRENT PRACTICE

Causal structure discovery is a research topic in many fields. Usually, causality is inferred from observing repeated co-occurrences of events. The Granger causality econometric is based on the principle that a causal event has unique information about its effects and this information cannot be found elsewhere. Implementations of Granger are built on simplistic assumptions, such as linearity, but still suffer from scalability and performance limitations. Biological sequencing approaches based on homology detection struggle to infer relationships in the presence of multiple co-occurring processes.

Perform scalable, near-real-time analysis to detect and identify event co-occurrences and assemble sequences of co-occurrences into recurrent temporal patterns.



An overview of our analytical framework for event analysis and recurrent pattern discovery.

U.S. DEPARTMENT OF
ENERGY

Recently, the Defense Advanced Research Projects Agency (DARPA) launched the Biochronicity Grand Challenge to make progress in this area of biological event modeling using temporal structure as a conserving factor. State of the art in cyber and network operations is divided between active and passive approaches. Active approaches monitor responses while perturbing (i.e., injecting faults into) the network. Passive approaches observe without modification. These approaches draw inferences by assuming that events are independent and identically distributed, which does not reflect the true nature of cyberspace.

TECHNICAL APPROACH

Our framework performs scalable, near-real-time analysis to (a) detect and identify event co-occurrences and (b) assemble sequences of co-occurrences into recurrent temporal patterns. The framework's core is an ensemble of machine learning, signal processing, and statistical methods to detect and identify recurrent temporal patterns in streaming data. The framework operates in three phases: event projection and transformation, event co-occurrence analysis, and recurrent pattern assembly.

In the first phase, data are transformed by categorizing them into classes using either unsupervised machine learning (appropriate for unstructured data or when the environment does not provide the target), supervised machine learning (trained on prior sets of data and when target responses are available), or expert- or rule-based categorizers. Each class represents an event type. Each class element (i.e., data tuple) is an instance of that event type. We call these instances "events." In the second phase, we perform event co-occurrence detection on system properties that are conserved. Domain- and problem-specific properties are conserved. Machine learning, signal processing, and statistical methods are then selected to recognize property conservation. In the third phase, we generalize over all event co-occurrences to model robust patterns of interactions.

Depending on the problem and domain, we may perform a temporal-only, spatial-only, or spatiotemporal-conserving analysis. To gain a higher abstraction of information, additional layers of machine learning may be appended to the aforementioned approach.

IMPACT

The ability to discover recurrent temporal patterns in a scalable and near-real-time fashion is fundamental to decomposing a mixture into its constituents. In cyberspace, such patterns characterize the causal relationship between cyber components. The patterns can be used to obtain data and graph size reduction, for example, by allowing only event sequences that do not match the patterns to be analyzed, transmitted to remote sites, or incorporated into graphs. Additional benefits include advances in privacy, preserving data mining as the patterns are generalizations of interactions.

Contacts

Thomas E. Carroll
Principal Investigator
(509) 371-6731
Thomas.Carroll@pnnl.gov

John R. Johnson
Program Director
(509) 375-2651
John.Johnson@pnnl.gov

Satish Chikkagoudar
Principal Investigator
(509) 375-2744
Satish.Chikkagoudar@pnnl.gov

Pacific Northwest
NATIONAL LABORATORY
Proudly Operated by **Battelle** *Since 1965*