**Invention Disclosure**

**Title**

Detecting Flow Anomalies in Distributed Systems

**Inventors**

Freddy Chua, Bernardo A. Huberman and Ee-Peng Lim

**Description**

We outline a method of detecting anomalies and their physical locations that lay hidden in the deep layers of networks within distributed systems. A functional network that transport information, physical entities or traffic is necessary for the operations of a distributed system.

In typical cases, the existence of anomalies within networks of distributed systems will cause severe disruptions, which draws the attention of its stakeholders such as users, owners and administrators of the distributed systems. Such anomalies are then easily detected and located for corrections.

However, our method focus on a specific type of *non-critical anomalies* that allows the distributed systems to continue their normal operations without significantly obstructing the distributed systems from meeting their objectives. Due to the fact that non-critical anomalies do not pose immediate threats, they are often ignored by administrators and owners of the distributed systems. If such non-critical anomalies are not corrected appropriately, either a long term physical deterioration or combination of several rare events may trigger a catastrophic failure in the distributed systems. It would be useful to utilize our method of anomalies detection in order to detect such non-critical anomalies early and prevent undesirable outcomes in the unforeseeable future.

An old, simple yet expensive method of performing such early detection is to install physical sensors in multiple locations of the networks within the distributed systems and monitor sensor readings for signs that are abnormal or unusual. We propose a cheap, non-intrusive, yet effective method to monitoring the flows of information, entities or traffic.

Our method performs the detection by utilizing data that is obtained from traffic flows which records the information such as resources used for completing the flow, or the amount of physical entities conserved after the flow. The non-intrusive constraint of our method requires us to only have information at the source and destination, while ignoring the detailed information during the flow. We could also assume that obtaining such detailed information during the flow requires expensive and intrusive installation of additional sensors within the networks.

The method comprised of two phases. The first phase uses all the available information in the data with assumptions from domain and contextual knowledge of the flow, to infer the missing information during the flow using a statistical model. The statistical model allows us to estimate the information we should observe at the destination of the flow, in terms of its mean and variance. By comparing the observation with the estimation (mean and variance), we would be able to know whether a flow is statistically deviated.

While the use of statistical deviation is common in anomalies detection work, a distinctive feature of our invention lies in the second phase, which addresses the insufficiency of statistical deviations as sole indicators of anomalies. In addition to the statistical deviation of each flow *f*, we also derive for flow *f*, the number of statistically deviated flows that are related to *f*. The definition of relation between a pair of flows would depend on the context and nature of the distributed system under investigation. In physical distributed systems, the relation would be defined in terms of the time and physical location of the flow. Using the defined relations, we would obtain an indication of whether a flow *f* is an anomaly, by positively correlating to the number of statistically deviated flows that are related to *f*. Using the end (source and destination) points of an anomalous flow, we would be able to isolate the physical location of the anomaly within the distributed system.

Please refer to Diagram as follows for an outline of the method.

Diagram

Find Relations between the Flows (Phase 2)

Statistical Model (Phase 1)

Input: Network Data of Distributed System

Output: Anomalous Flows and Location of Anomalies