Event Based Community Detection for Networks

Patrick A. O'Neil
Michael D. Porter

GeoEye Analytics

May 1, 2012

GeoEye Analytics

Problem Description

Motivation

Given a dynamic network and a set of events for which the network is known to be responsible, it is natural to ask questions about which nodes participated in the events. Uncovering this information reveals details about the network's activity, such as which nodes are most responsible for the network's past activity.

Problem Description

Motivation

Given a dynamic network and a set of events for which the network is known to be responsible, it is natural to ask questions about which nodes participated in the events. Uncovering this information reveals details about the network's activity, such as which nodes are most responsible for the network's past activity.

Objective

Given a dynamic network and a set of events, for each node, we would like to determine a subset of events in which that node participated.

Assumptions

- Our primary assumption is that nodes who are involved with an event will have an anomalous neighborhood network structure around the time of the event.
- The event set will be sparse (i.e. there will be few events).
- Nodes who have worked together in the past will likely work together again at some point in the future.
- A node's usual behavior remains relatively constant during the course of observation.

GeoEye Analytics

Network & Event Notation

- Let $G_t(V, E)$ be a weighted graph at time $t \in \{1, 2, ..., T\}$ with a set of nodes V and edges E.
- Let $w_t(\{v_1, v_2\}) \in \mathbb{N}$ denote the weight of the edge between nodes $v_1, v_2 \in V$ at time t, 0 if nodes v_1 and v_2 are not actually connected.
- For $v \in V$ let $N_t(v)$ be the set of neighbors of v and $E_t(v)$ be the set of edges connected to v at time t.
- Let $A = \{a_1, a_2, ..., a_{|A|}\}$ be an event set where a_i denotes the time of event i.

Structural-EPD

Structural Event-Participation Detection

Seeks to find anomalous neighborhood structure by looking for times when a node either changed who it was communicating with or the frequency with which it was communicating with other nodes.

Thus, for node v, we are looking for anomalies in the set $N_t(v)$ and/or the set $\{w_t(v,u) : u \in V(G)\}$ for t near event times.

Methods for S-EPD

There are many ways to model the communication of a node's neighborhood. Two methods will be discussed here.

- Counting Process for each potential edge
- Distance from Median Graph

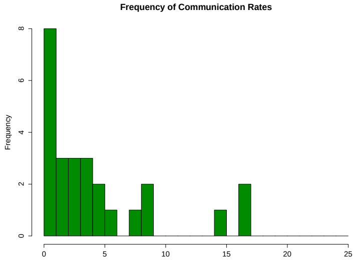
Counting Process

- This approach models the communication between a pair of nodes *during non-event times* as a counting process.
- Since most nodes do not communicate with each other, we will employ a hurdle model.
- For nodes u and v, let $C_t(u, v)$ be the number of times u and v communicated during time t.
- We model $C_t(u, v) = 0$ and $C_t(u, v) > 0$ using a binomial distribution.
- For $C_t(u, v) > 0$, we model C using a geometric distribution setting

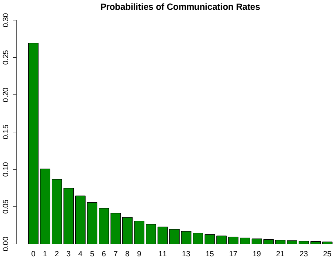
$$p = \frac{1}{1 + E[C_s(u, v)]} \text{ with } s \in \{t : C_t(u, v) > 0\}$$

Counting Process

Below is an example of the counting process model for communication between two nodes.



(a) Realized



(b) Hurdle Model

Counting Process

- For node u, let $C_t(u) = \{c_{v_1}, c_{v_2}, ..., c_{v_k}\}$ represent the number of times u communicated with each $v_i \in V$ at time t.
- For each c_{v_i}, we calculate $P(C_t(u, v_i) = c_{v_i})$, the probability that u communicates with node v_i c_{v_i} times.
- Assuming communication rates from node to node are independent, we find the joint probability $P(C(u) = C_t(u)) = \prod P(C_t(u, v_i))$, the probablity that this communication structure would occur.
- Unusually low probabilities are considered indicative of anomalous neighborhood network structure.

S-EPD: Distance from Median Graph

Definition: Edit Distance

Given two graphs G and G', each with the same number of vertices, the edit distance $D : G \times G \to \mathbb{N}$ between G and G' is defined as $D(G, G') = |E(G) \triangle E(G')|$.

GeoEye Analytics

S-EPD: Distance from Median Graph

Definition: Edit Distance

Given two graphs G and G', each with the same number of vertices, the edit distance $D : G \times G \rightarrow \mathbb{N}$ between G and G' is defined as $D(G, G') = |E(G) \triangle E(G')|$.

Definition: Median Graph

The median graph $\overline{G}_H$ of a set of graphs $H = \{G_1, G_2, ..., G_m\}$ each with n vertices is defined as,

$$\overline{G}_H = \underset{G \in \mathbb{G}_n}{\operatorname{argmin}} \sum_{G_i \in H} D(G, G_i)$$

where $\mathbb{G}_n$ is the set of all graphs constructible from n vertices.

S-EPD: Distance from Median Graph

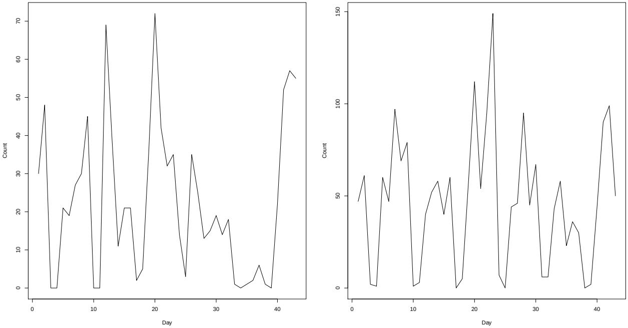
Framed for our problem,

- Let H be the set of graphs during which events did not occur. We first calculate the median graph, $\overline{G}_H$, of H.
- Then for every graph G_t with $t \in \{1, 2, ..., T\}$, we calculate $D(G_t, \overline{G}_H)$, the edit-distance between the graph and the median graph.
- Times with significantly large edit-distances are considered anomalous. *We search for nodes which exhibit anomalous neighborhood structure around the time of events.*

GeoEye Analytics

S-EPD Example

In this example, the plots show the communication rates of two nodes. The node on the left was involved with an activity (going on vacation) around times 32-38 while the node at the right acted normally during the period of interest.



(a) Participant (b) Non-Participant

GeoEye Analytics

Metric-EPD

Metric Event-Participation Detection

While structural EPD examines the communication behavior of a particular node, metric EPD determines how the role of a node changes through time. Using SNA metrics, we can look for anomalous positioning in the network as well as local neighborhood structure.

A variety of multivariate time-series anomaly detection methods exist and can be utilized for M-EPD.

GeoEye Analytics

1. Preliminaries
   - Problem Description
   - Assumptions
   - Network & Event Notation
2. Event Participation Detection
   - Structural
   - Metric-EPD
3. **Tie-Strength Clustering**
   - **Tie-Strength**
   - **Clustering**
4. Network Activity Score
   - NAS Model
   - NAS Prediction
5. Results

GeoEye Analytics

Tie-Strength Metrics

- Given a set of network members N and a set of events A, we can construct a bipartite graph $EP = G(V, E)$ with $V \subseteq N \cup A$ and $E \subseteq N \times A$.
- An edge exists between a network member and an event when the network member is believed to have participated in that event.
- For tie-strength, we use the Adamic & Adar tie-strength metric,

$$TS(u, v) = \sum_{e \in \Gamma(u) \cap \Gamma(v)} \frac{1}{\log |\Gamma(e)|},$$

where $\Gamma(u)$ is the neighborhood of node u (i.e. the events in which u participated).

GeoEye Analytics

Event-Based Clustering

- We construct a weighted graph G_{TS} where the nodes are the members of the network and where the weight of an edge $\{v_1, v_2\}$ of G_{TS} is the tie-strength between v_1, v_2.
- Running a clustering algorithm on this weighted graph produces a list of clusters of nodes who participated in the same events.

GeoEye Analytics

Network Activity Score Model

So far we have the following;

- Anomaly scores for each node at each time period.
- For each event, a list of nodes that are predicted to have participated in that event.
- Clusters of nodes that work together.

The obvious next step is to track how anomalous these clusters are behaving in the hope of predicting when the cluster might produce another event.

Network Activity Score Model

Cluster Anomaly Scores

For each cluster i, we aggregate the anomaly scores of the involved nodes to get a cluster anomaly score $CS_i(\mathbf{y})$ where $\mathbf{y}$ is the set of anomaly scores of each node in cluster i.

Network Anomaly Score

Then for the Network Activity Score, we aggregate the cluster anomaly scores to obtain the Network Activity Score, $NS(\mathbf{z})$, where $\mathbf{z}$ are the cluster scores.

Network Activity Prediction

Tracking these scores over time will hopefully give us an indication of when future events might occur (i.e. some important clusters are beginning to act anomalously).

1. Preliminaries
   - Problem Description
   - Assumptions
   - Network & Event Notation
2. Event Participation Detection
   - Structural
   - Metric-EPD
3. Tie-Strength Clustering
   - Tie-Strength
   - Clustering
4. Network Activity Score
   - NAS Model
   - NAS Prediction
5. Results

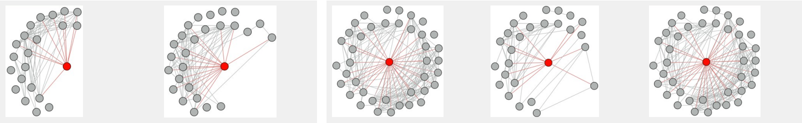
GeoEye Analytics

Preliminary Results: DCNS

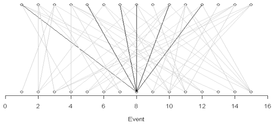
Dynamic, Covert Network Simulation

- DCNS is a covert network simulation tool which seeks to mimic real world covert networks.
- The network seeks to remain secretive while accomplishing various objectives.
- The network is composed of “cells” which carry out the tasks (aquisition of resources, attacks, etc).
- There are external interventions (members captured/killed) and the network responds to these interventions by changing its structure.

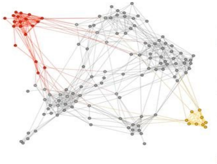
Preliminary Results: DCNS



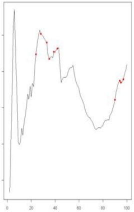
(a) EPD



(b) Participation Graph



(c) Event-Based Clustering



(d) Network Activity Score

GeoEye Analytics

Preliminary Results: DCNS

The following shows the percent of event based communities who were actually involved in the same events. Each cluster had around 10 members.