Event Based Community Detection for Networks

Patrick A. O'Neil

GeoEye Analytics

December 4, 2012

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.

Network & Event Notation

- Let $G_t = (V, E_t)$ be a weighted, directed graph at time $t \in \{1, 2, ..., T\}$ with a set of nodes V and weighted edges E_t.
- Let $w_t(v_1, v_2) \in \mathbb{N}$ denote the weight of the edge from node $v_1 \in V$ to $v_2 \in V$ at time t. If there is no such edge, $w_t(v_1, v_2) = 0$.
- For $v \in V$ let $\Gamma_t(v)$ be the set of neighbors of v and $E_t(v)$ be the set of edges connected to v at time t.
- Let $A = \{a_1, a_2, ..., a_k\}$ be an event set where a_i denotes the time of event i and $1 \leq a_1 < ... < a_i < ... < a_k \leq T$.

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 $\Gamma_t(v)$ and/or the sets $\{w_t(v,u) : u \in V(G)\}$ and $\{w_t(u,v) : u \in V(G)\}$ 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

S-EPD: Counting Process

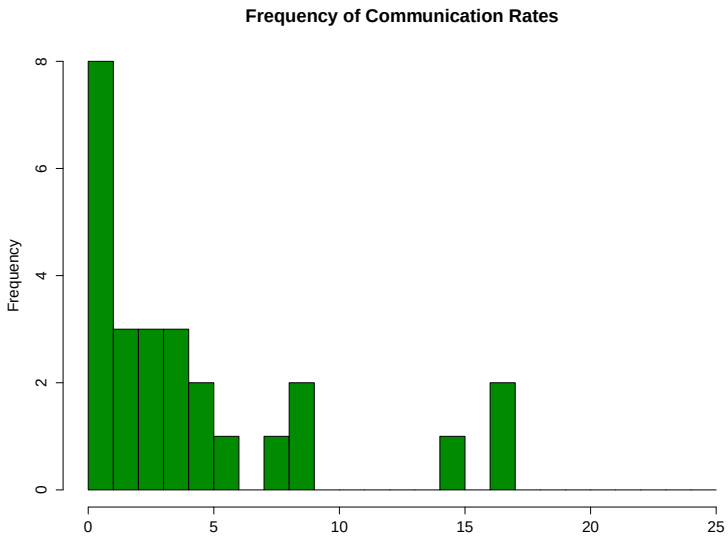
- This approach models the communication between a pair of nodes *during non-event times* as a counting process.
- Since, at any given time, most nodes do not communicate with each other, we will employ a hurdle model.
- We model $w_t(u, v) = 0$ and $w_t(u, v) > 0$ using a binomial distribution (similarly for $w_t(v, u)$).
- For $w_t(u, v) > 0$, we model $w_t(u, v)$ using a shifted geometric distribution with

$$p = 1 - \frac{E[w_s(u, v)]}{1 + E[w_s(u, v)]} \text{ with } s \in \{t : w_t(u, v) > 0, t \notin A\}$$

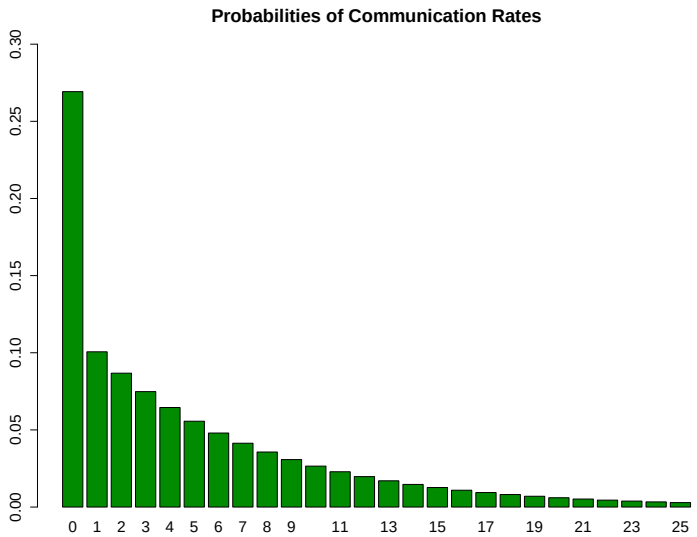
$$P(w_t(u, v) = k) = \left(\frac{E[w_s(u, v)]}{1 + E[w_s(u, v)]}\right)^{k-1} \left(1 - \frac{E[w_s(u, v)]}{1 + E[w_s(u, v)]}\right)$$

S-EPD: Counting Process

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



(a) Realized



(b) Hurdle Model

S-EPD: Counting Process

- For node u, let $C_t : V \to \mathbb{N}^{n-1}$ be the map which gives the number of times u communicated with each $v_i \in V$ at time t (i.e. $C_t(u) = \{c_{v_1}, c_{v_2}, ..., c_{v_k}\}$ where c_{v_i} is the number of times u communicated with v_i at time t)
- For each c_{v_i}, we calculate $P(w_t(u, v_i) = c_{v_i})$, 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 $P(C(u) = C_t(u)) = \prod P(w_t(u, v_i) = c_{v_i})$, 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 $D(G, G') = |E(G) \triangle E(G')| = |\,(E(G) \backslash E(G')) \bigcup (E(G') \backslash E(G))\,|$.

Note that,

$$\begin{aligned} D(G_1, G_3) &= |E(G_1) \triangle E(G_3)| = |\,[E(G_1) \triangle E(G_2)] \triangle [E(G_2) \triangle E(G_3)]\,| \\ &\leq |E(G_1) \triangle E(G_2)| + |E(G_2) \triangle E(G_3)| = D(G_1, G_2) + D(G_2, G_3) \end{aligned}$$

Therefore, D is a metric (the other two conditions are obvious) and the space of graphs can be considered a metric space.

S-EPD: Distance from Median Graph

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.

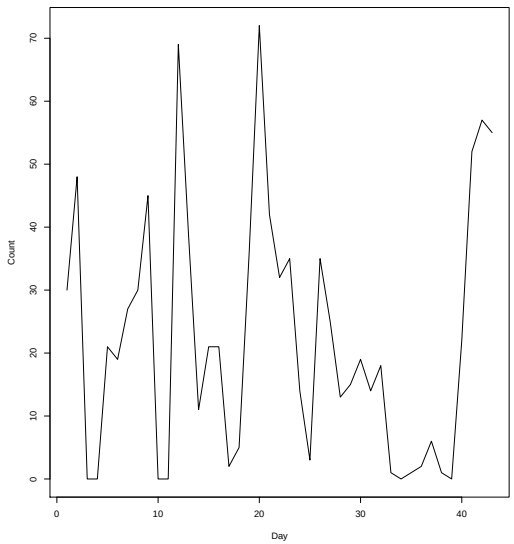
S-EPD: Distance from Median Graph

Framed for our problem,

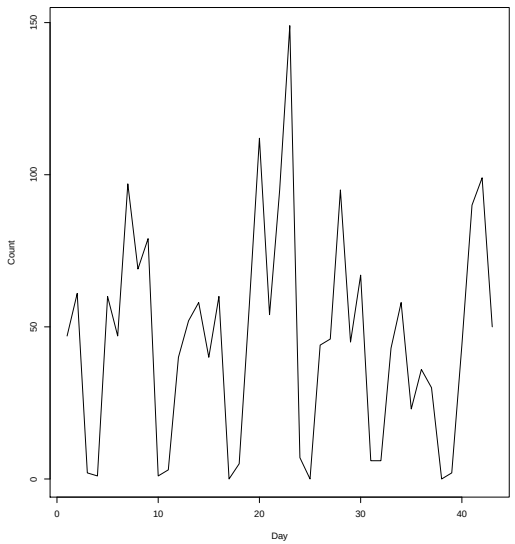
- Let H be the set of graphs during which events did not occur. We first calculate the median graph, $\overline{G}_H$, of H.
- Then for every graph G_t with $t \in \{1, 2, ..., T\}$, we calculate $D(G_t, \overline{G}_H)$, the edit-distance between the graph and the median graph.
- Times with significantly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

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. Note that the total communication rate remained relatively constant (this employee continued working while on vacation), but the rates between individuals changes.



(c) Participant



(d) Non-Participant

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.

Methods for M-EPD

Example SNA metrics:

- Anomalous tie-strength (Adamic & Adar):

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

- Anomalous Jaccard Index:

$$J(u,v) = \frac{|\Gamma(u) \cap \Gamma(v)|}{|\Gamma(u) \cup \Gamma(v)|}$$

- Anomalous betweenness centrality,

$$g(v) = \sum_{u \neq v \neq w} \frac{\sigma_{uw}(v)}{\sigma_{uw}}$$

where $\sigma_{uw}(v)$ is the number of shortest paths from u to w which pass through v and σ_{uw} is the number of shortest paths from u to w.

Tie-Strength Metrics

- Given a set of network members N and a set of events A, we can construct a bipartite graph $EP = (V, E)$ 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).

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.
- I will give a brief overview of the *Shrink-H* clustering algorithm.

Shrink Clustering

Modularity (Q): the fraction of the edges that fall within the given groups minus the expected such fraction if edges were distributed at random.

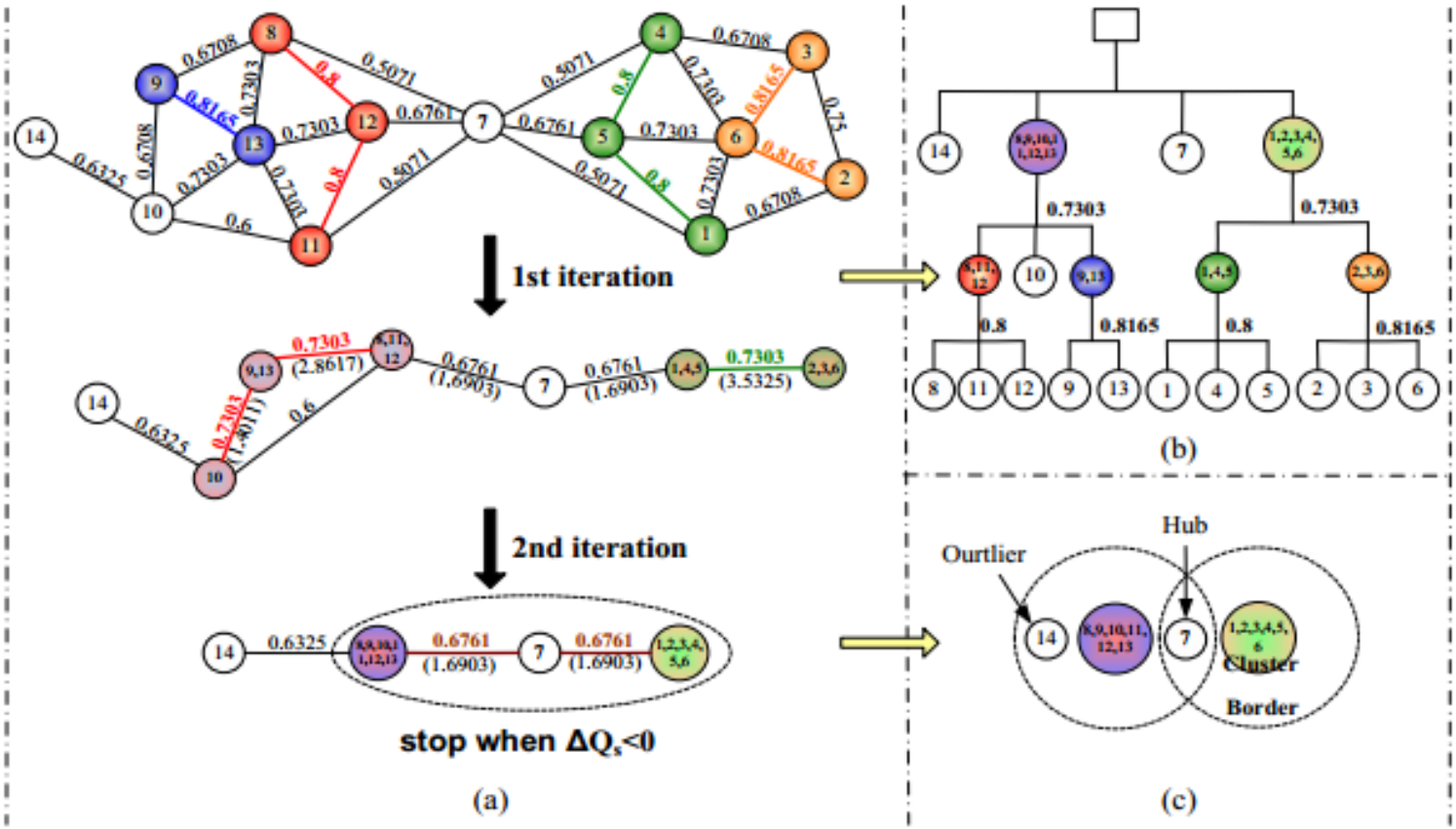
Structural Similarity

Given a graph $G = (V, E)$, let $\Gamma(u)$ denote the neighbhorhood of $u \in V$. Then the *structural similarity* $\gamma(u, v)$ of two nodes $u, v \in V$ is given by,

$$\gamma(u, v) = \frac{\sum_{x \in \Gamma(u) \cap \Gamma(v)} w(u, x) w(v, x)}{\sqrt{\sum_{x \in \Gamma(u)} w^2(u, x)} \sqrt{\sum_{x \in \Gamma(v)} w^2(v, x)}}$$

The Shrink-H algorithm combines nodes with high structural similarity into "supernodes" (and in turn combines supernodes with high structural similarity) until the modularity gain ΔQ becomes negative.

Shrink Clustering: A Picture



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 using a logistic regression model,

$$CS_i(y) = \frac{e^y}{e^y + 1},$$

where $y = \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + ... + \alpha_n x_n$ and $x_1, ..., x_n$ are the anomaly scores for the nodes in the cluster.

In this framework, α_j represents how much node j's behavior impacts the cluster's event rate.

Network Activity Score Model

Network Anomaly Scores

Then for the Network Activity Score, we perform another logistic regression with all the events as the dependent variable,

$$NS(z) = \frac{e^z}{e^z + 1},$$

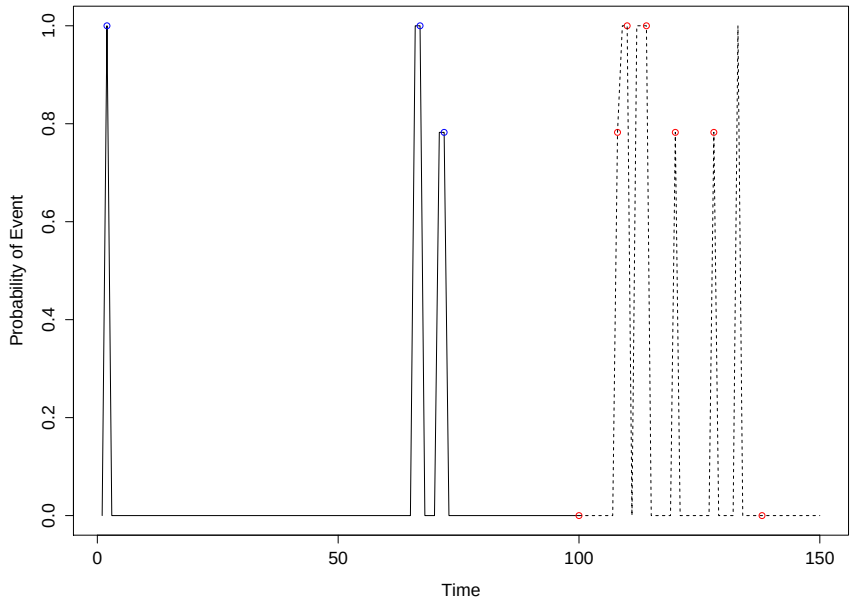
where $z = \beta_0 + \beta_1 CS_1 + \beta_2 CS_2 + ... + \beta_m CS_m$ for cluster anomaly scores $CS_1, CS_2, ..., CS_m$.

In this framework, β_i represents how much cluster i's behavior impacts the network's event rate.

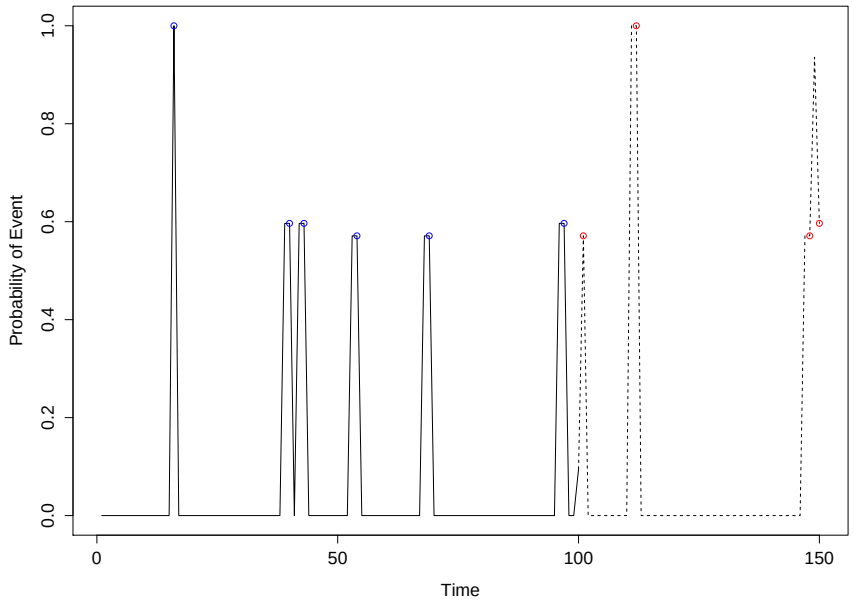
Network Activity Prediction

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).

Results: DCNS

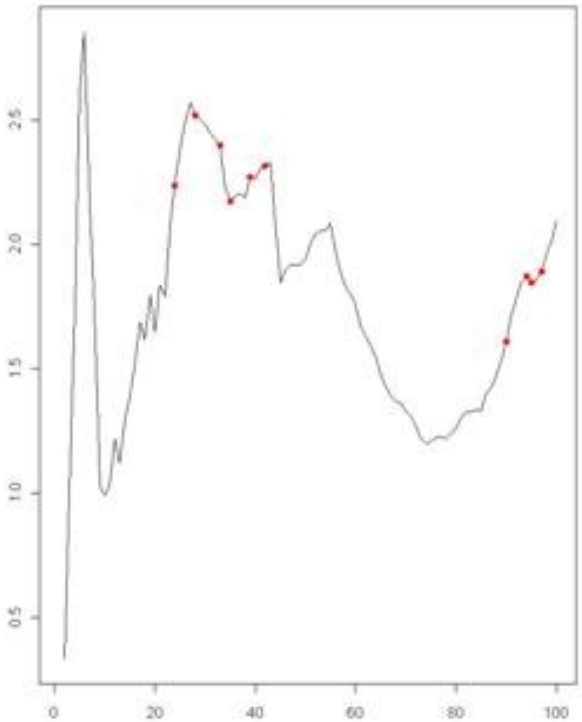


(e) Run 1



(f) Run 2

Results: GeoEye Email Network



References

1. Huang et al, *SHRINK: A Structural Clustering Algorithm for Detecting Hierarchical Communities in Networks*, CIKM '10, 2010
2. Gupte & Eliassi-Rad, *An Axiomatic Approach to Tie-Strength Measures*, NIPS 2011