Inference in time series of graphs using locality statistics

Heng Wang, Minh Tang, Youngser Park, Carey E. Priebe
Johns Hopkins University
Department of Applied Mathematics and Statistics
Baltimore, Maryland 21218-2682 USA
hwang82@jhu.edu, mtang10@jhu.edu, youngser@jhu.edu, cep@jhu.edu

Abstract

We formulate change-point detection in a time series of graphs as a hypothesis testing problem in terms of Stochastic Block Model time series. We analyze two classes of scan statistics by deriving the limiting properties and power characteristics of the competing scan statistics.

1. Change-Point detection in Stochastic Block Model formulation

Given a time series of graphs $G_t = (V_t, E_t)$, where the vertex set $V = [n] = \{1, \ldots, n\}$ is fixed throughout, an important inference task in time series analysis is to identify, from $(G_t)$, excessive communication activities in a subregion of a dynamic network. Statistically speaking, we test, for a given $t \in \mathbb{N}$, the null hypothesis $H_0$ that $t$ is not a change-point against the alternative hypothesis $H_1$ that $t$ is a change-point.

We say that $t$ is a change-point for $(G_t)$ if there exists distinct choices of matrices $P^0$, $P^1$ independent of $t$ such that

$H_0: G_t \sim \text{SBBM}(P^0, [n])$ for all $t$, \quad $H_1: G_t \sim \text{SBBM}(P^1, [n])$ for $t \leq t^* - 1$

where $\text{SBBM}(P, [n])$ denotes the stochastic blockmodel of $[1]$, with block connectivity probabilities $P$ and unknown block memberships $(v_i)$. In each block $[u]$, vertices follow the same probabilistic behavior and $P$ is a $d \times d$ symmetric matrix where $P_{ij}$ denotes the block connectivity probability between blocks $i$ and $j$. In this work, we illustrate a subset of vertices with chatty anomalous behavior in an otherwise stationary setting, we are concerned about is of the particular form for some $k > 0$,

$P^0 = \begin{pmatrix} p_{11} & \cdots & p_{1n} \\ \vdots & \ddots & \vdots \\ p_{n1} & \cdots & p_{nn} \end{pmatrix}$, \quad $P^1 = \begin{pmatrix} p_{11} & \cdots & p_{1n} \\ \vdots & \ddots & \vdots \\ p_{n1} & \cdots & p_{nn} \end{pmatrix}
$

The case where $p_{12}, \ldots, p_{1n} > p_{i+1} > p_{i}$ is of interest because we can consider each of the $[u]$ as representing a "chatty" group for time $t \leq t^* - 1$, and at $t^*$, the previously non-chatty group $[u]$ becomes more chatty. The detection of this transition for the vertices in $[u]$ is one of the main reasons behind the locality statistics that will be explored.

Figure 1: Notional depiction of a 3-block time series of graphs in which the anomaly occurs at time $t^*$, a subset of vertices exhibits a change in behavior. When testing for change at time $t^*$, the recent past graphs $G_{t^* - 1}, G_{t^* - 2}, \ldots, G_{t^* - k}$ are used to standardize the invariants.

2. Locality Statistics and Graph Invariants

2.1 Two locality statistics

Let $N_t(v, G_t) = \{u \in V | (u, v) \in E_t\} \leq k$ and $N_t'(G_t)$ denote the subgraph of $G_t$ induced by $V_t'$. We now define two different but related locality statistics on $(G_t)$. For a given $t$, let $\Psi_{t}(v)$, introduced in [2], be defined for all $k \geq 1$ and $v \in V$ by

$\Psi_{t}(v) = |E_t(N_t(v, G_t))|$. \quad (1)

We define another related locality statistic, $\phi_{t}(v)$, introduced in [3], for all $k \geq 1$ and $v \in V$ by

$\phi_{t}(v) = |E_t(N_t(v, G_t))|$. \quad (2)

$\phi_{t}(v)$ counts the number of edges in the subgraph of $G_t$ induced by $N_t(v, G_t)$. Let $t$ and $t'$ be given, with $t' \leq t$.

2.2 Temporally-normalized statistics

Let $\Delta_{t,k}$ be either the locality statistic $\phi_{t}(v)$ or $\phi_{t}(v)$, where for ease of exposition the index $i$ is a dummy index when $\phi_{t}(v) = \phi_{t}(v)$. With the purpose of determining whether $t$ is a change-point, we now define two normalized statistics $\Delta_{t,k}(v)$ for $t 

\Delta_{t,k}(v) = \frac{\Delta_{t,k}(v)}{\max(\Delta_{1:k}(v))}$ \quad (3)

We then consider the maximum of these vertex-dependent normalizations $\max_{t \in [1:k]}(\Delta_{t,k}(v))$ and refer to $\max_{t \in [1:k]}(\Delta_{t,k}(v))$ as the localized scan statistic.

Finally, for a given integer $t \geq 0$ and $v \in V$, we define the temporal normalization of $M_{t,k}(v)$ by

$S_{t,k}(v) = \frac{M_{t,k}(v)}{\max_{t \in [1:k]}(M_{t,k}(v))}$. \quad (4)

3. Limiting Theory and Experiment

Theorem Under both $H_0$ and $H_1$, the limiting $S_{t,k}(\Theta)$, $S_{t,k}(\Phi)$, $S_{t,k}(\phi_{t})$, and $S_{t,k}(\phi_{t})$ are the maxima of random variables which, under proper normalizations, follow a standard Gumbel $(0, 1)$ distribution in the limit.

Figure 2: An example to illustrate the calculation of $\Delta_{t,k}(v)$ with varying underlying statistics ($\Psi_{t}(v)$ or $\phi_{t}(v)$) and order distances $(d = 0 \lor k = 1)$. \quad \begin{pmatrix} \Phi_{t} & \Psi_{t} \\ 5 & 7 \\ 2 & 4 \\ 3 & 2 \\ 1 & 3 \end{pmatrix}

Figure 3: A comparison, using the limiting properties of $S_{t,k}(\Theta)$ and $S_{t,k}(\Phi)$, of $\Delta_{t,k}(\Phi)$ for $\Delta_{t,k}(\Phi)$ and $\Delta_{t,k}(\Theta)$, for different null and alternative hypotheses par as normalized by $\Phi$ and $\Theta$, respectively. The blue-colored region correspond to values of $\Delta_{t,k}(\Phi)$ for which $\Delta_{t,k}(\Phi) > \Delta_{t,k}(\Theta)$, while the red-colored region correspond to values of $\Delta_{t,k}(\Phi)$ for which $\Delta_{t,k}(\Theta) > \Delta_{t,k}(\Theta)$.

4. Future Work

Locality statistics based on $\Theta$ can be readily computed in a real-time streaming data environment, in contrast to those based on $\Phi$. Thus, discovering approximations of locality statistics based on $\Theta$ which simultaneously maintain better power characteristics and are amenable to streaming graphs is of interest.

References