APDL: A Probabilistic Modeling Language for Anomalous Pattern Detection on Large Graphs

Feng Chen
Department of Computer Science
University at Albany, SUNY

Monday 15th June, 2015

Anomalous Pattren Detection (APD) detection in graph data is a ubiquitous task with a wide variety of real world applications including disease surveillance (where we must detect emerging disease outbreaks at a very early stage), network intrusion detection (where we attempt to identify patterns of suspicious network activity) and road traffic congestion detection (where we attempt to detect regions of non-recurrent congestion). Although numerous methods have been proposed to tackle this problem in different scenarios, most of these methods can be framed in a unified optimization framework. We consider a graph $G = (\mathbb{V}, \mathbb{E}, x)$, where $\mathbb{V}$ refers to the set of vertices, $\mathbb{E} \subseteq \mathbb{V} \times \mathbb{V}$ refers to the set of edges, and $x(v) : v \to \mathbb{R}^D$ maps each vertex v to a vector of D features. The unified optimization framework is formalized as:

$$\max_{S \in \mathscr{B}(\mathbb{V})} F(S) + \lambda \Phi(S), \tag{1}$$

where $\mathscr{B}(\mathbb{V}) \subseteq \mathscr{P}(\mathbb{V})$ refers to the space of feasible subsets of vertices, i.e., $\mathscr{B}(\mathbb{V}) = \{S | S \subseteq \mathbb{V}, S \text{ is connected}\}$, $\mathscr{P}(\mathbb{V})$ refers to the power set of $\mathbb{V}$; $F(S)$ refers to a utility funciton measuring the degree of interestingness or anomalousness of the subset S, i.e., $F(S) = \sum_{v \in S} \sum_{i=1}^{D} x^i(v)$; $\Phi(S)$ refers to a regularization component that is usually used to characterize user's prior knowledge and special requirements about the subset S, i.e., $\Phi(S) = |S \cap \{v_0\}|$, where v_0 is known to belong to S; and $\lambda \in [0, \infty]$ is the tradeoff parameter.

Depending on the specification of the function $F(S)$, existing methods can be classified to two categories, namely, parametric and nonparametric methods. Parametric methods assume specific forms of distributions for features of normal and abnormal vertices respectively, and formalize anomaly detection as a hypothesis testing problem. In particular, under the alternative hypothesis ($H_1(S)$), an underlying anomalous phenomenon is characterized by the following: features of a majority of the vertices are generated from the same background distribution, and features of perhaps a small connected subset $S \subseteq \mathbb{V}$ of vertices are generated from a different distribution. The goal is to maximize an appropriate set function ($F(S)$), typically the likelihood ratio $F(S) = \frac{\Pr(\text{Data}|H_1(S))}{\Pr(\text{Data}|H_0)}$, over all possible connected subsets S (with H_0 being the null hypothesis). Depending on specific forms of distributions assumed, a number of methods have been proposed, including expectation-based Poisson statistic [9], Kulldorff statistic [6], elevated mean scan statistic [12, 10], and various others.

Nonparametric methods do not assume specific forms of distributions for normal and abnormal vertices. Instead, they first estimate a p-value for each vertex based on empirical calibration by comparing the current features of this vertex with its features in the historical data for the vertex [4, 7]. The empirical p-value provides an estimate of the probability that a randomly selected sample would have observed features as extreme as the current features of this vertex, under the null hypothesis that no events of interest are occurring. This approach then maximizes a score function $F(S)$ of p-values in S, typically nonparametric scan statistic measuring the significance of the collection of p-values in S, over all possible connected subsets. A number of NPGS statistic functions have been proposed in recent years, including Berk-Jones (BJ) statistic [2], Higher Criticism (HC) statistic [5], Tippet's statistic, rank truncated statistic, and various others. Note that, these nonparametric statistic functions were originally proposed to combine p-values from a set of hypothesis tests in the area of statistical meta analysis. Recent studies show that these functions can be well applied to NPGS for detecting anomalous subgraphs [8, 4, 3].

A number of structure constraints on the subset S are considered to define $\mathscr{B}(\mathbb{V})$. The general constraint of connectivity on S is the most popular one [9, 6, 12, 10, 4, 7]: $\mathscr{B}(\mathbb{V}) = \{S | S \subseteq \mathbb{V}, S \text{ is connected}\}$. However, the

Table 1: APDL language Defined Syntax

Syntax Name	example	Description
Data type	real, int ,set	APDML defined main data type
Constrain	real, int: "!=","$>$", "$\geq$","$<$","$\leq$"	Number constrains
	Set:in, not in, is connected	Set theory constrains
Hypothesis	"null hypothesis", "alternative hypothesis"	
if ... else...	if statement	
for ... in ...:	for loop	

size of $\mathscr{B}(\mathbb{V})$ is exponential to the size of $\mathbb{V}$, and the resulting APD problem is hard to solve in general. In order to reduce the search pace $\mathscr{B}(\mathbb{V})$, a cardinality constraint of size K is considered: $\mathscr{B}_K(\mathbb{V}) = \{S|S \in \mathscr{B}(\mathbb{V}), |S| \leq K\}$. Spatial shapes, such as circle and rectangle, are considered when the graph of interest is a spatial graph (e.g., city-city network, transportation network) [1], and the cardinality of $\mathscr{B}(\mathbb{V})$ is quadratic to the cardinality of $\mathbb{V}$ for these shapes. For general graphs, compactness are proposed to approximate connectivity. A number of definitions of compactness of S have been proposed, including: 1) the sum of pair of distances in S [11]; 2) Steiner tree cost [11]; 3) edge-lasso regularization [13]; and 4) graph Laplacian regularization [14]. The compactness is used as the regularization component $\Phi(S)$, as it is difficult to define the space of compacted subsets using $\mathscr{B}(\mathbb{V})$.

Although a large number of methods have been proposed for APD, each method is designed based on specific assumptions and is not applicable when some of the assumptions are violated. To the best of our knowledge, there is no existing framework that supports a variety of anomalous patterns as defined above and allows users to incorporate prior knowledge and special constraints. This paper presents APDL, a probabilistic modeling language for detecting anomalous patterns in large graphs.

The basic symbols of the APDL language are shown in Table 1 and introduced using the example of disease surveillance. Suppose data $\equiv \{d_1, \cdots, d_N\}$, where $d_i \equiv (c_i^t, b_i^t)$, c_i^t refers to the number of reported respiratory cases in a county s_i on day t, and b_i^t refers to the expected count calculated based on the historical data. We assume a Possion distribution $c_i^t \sim \text{Possion}(b_i^t)$ for normal data records, and a different Poisson distribution $c_i^t \sim \text{Possion}(q \cdot b_i^t)$ for anomalous data records, where q, $q > 1$, is a unknown parameter that can be estimated via maximum likelihood estimation (MLE). Then we obtain expectation-based Poisson (EP) statistic, and the log-likelihood ratio can be derived as $F(S) = C\log(C/B) + B - C$, if $C > B$, and $F(S) = 0$ otherwise, where C and B are respectively the aggregate count $\sum_{i \in S} c_i^t$ and aggregate baseline $\sum_{i \in S} b_i^t$. The anomalous pattern defined above can be described in the APDL language as shown in Figure 1 (a). Given the file name "FileName", where the set of vertices $\mathbb{V}$, the set of edges $\mathbb{E}$, and the counts and bases can be loaded, the compiling of APDL will generate the anomalous subset S of interest, which has the largest expectation-based Poisson statistic score over all connected subsets in the input graph.

We consider the second well-known example, namely, Kulldorff's statistic. Under the null hypothesis ($H_0(S)$), we assume a Poission distribution: $c_i^t \sim \text{Poisson}(q_{all} \cdot b_i^t)$, where q_{all} is an unknown parameter; under the alternative hypothesis ($H_1(S)$), we assume different Poisson distributions for normal and abnormal records, respectively: $c_i^t \sim \text{Poisson}(q_{in} \cdot b_i^t)$, if $v_i \in S$; $c_i^t \sim \text{Poisson}(q_{out} \cdot b_i^t)$, otherwise. There are three parameters in total, including q_{all}, q_{in}, and q_{out}. The Kulldorff's scan statistic function is defined as $F(S) = C(S)\log\frac{C(S)}{B(S)} + (C_{all} - C(S))\log(\frac{C_{all}-C(S)}{B_{all}-B(S)}) - C_{all}\log(\frac{C_{all}}{B_{all}})$, where $C(S) = \sum_{s_i^t \in S} c_i^t$, $C_{all} = C(\mathbb{V})$, $B(S) = \sum_{s_i \in S} b_i^t$, and $B_{all} = B(\mathbb{V})$. The anomalous pattern characterized by the Kulldorff's scan statistic can be simply described in APDL as shown in Figure 1 (c).

Figure 1 (b) demonstrates a more complicated version of expectation-based Poisson statistic, with two additional constraints on S. The first constraint requires the subset S to contain all the vertices in $S_0 = \{1,3,5\}$, where each number refers to the index number of a specific vertex. The second constraint requires the subset S to contain at most one vertex in $S_1 = \{2,4\}$, which is similar to can-not-link constraint in supervised clustering [1]. The optimization algorithms of the expectation-based Poisson statistic without and with these two constraints are different. However, users just need to describe the anomalous pattern detection problem using the APDL language, and do not need to care about the detailed implementation.

```
real q
constrain(q > 1)
V,E,C,B = LoadGraphData(FileName)
set S
constrain(S ⊆ V)
constrain(S is connected)
constrain(S is connected)
hypothesis = {null, alternative}
if hypothesis == null:
  for v in V:
    C(v) ~ Poisson(B(v))
else hypothesis == alternative:
  for v in S:
    C(v) ~ Poisson(q * B(v))
  for v not in S:
    C(v) ~ Poisson(B(v))
Infer S
```

(a) EP Statistic

```
real q
constrain(q > 1)
V,E,C,B = LoadGraphData(FileName)
set S
constrain(S ⊆ V)
constrain(S is connected)
set S0 = {1,3,5}
constrain(S0 ⊆ S)
set S0 = {2,4}
constrain(|S0 ∩ S| ≤ 1)
hypothesis = {null, alternative}
if hypothesis == null:
  for v in V:
    C(v) ~ Poisson(B(v))
else hypothesis == alternative:
  for v in S:
    C(v) ~ Poisson(q * B(v))
  for v not in S:
    C(v) ~ Poisson(B(v))
Infer S
```

(b) EP Statistic + 2 Constraints

```
real q_all
real q_in
real q_out
constrain(q_all > 0)
constrain(q_in > q_out)
constrain(q_out > 0)
V,E,C,B = LoadGraphData(FileName)
set S
constrain(S ⊆ V)
hypothesis = {null, alternative}
if hypothesis == null:
  for v in V:
    C(v) ~ Poisson(q_all * B(v))
else hypothesis == alternative:
  for v in S:
    C(v) ~ Poisson(q_in * B(v))
  for v not in S:
    C(v) ~ Poisson(q_out * B(v))
Infer S
```

(c) Kulldorff's statistic

Figure 1: Examples of APDL Language

References

[1] S. Basu, M. Bilenko, and R. J. Mooney. A probabilistic framework for semi-supervised clustering. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, pages 59–68, New York, NY, USA, 2004. ACM.

[2] R. H. Berk and D. H. Jones. Goodness-of-fit test statistics that dominate the kolmogorov statistics. *Z. Wahrsch. Verw. Gebiete*, 47(1):4759, 1979.

[3] P. Bogdanov, M. Mongiov, and A. K. Singh. Mining heavy subgraphs in time-evolving networks. In D. J. Cook, J. Pei, W. W. 0010, O. R. Zaane, and X. Wu, editors, *ICDM*, pages 81–90. IEEE Computer Society, 2011.

[4] F. Chen and D. B. Neill. Non-parametric scan statistics for event detection and forecasting in heterogeneous social media graphs. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, pages 1166–1175, 2014.

[5] D. Donoho and J. Jin. Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.*, 32(3):962–994, 06 2004.

[6] M. Kulldorff. A spatial scan statistic. *Communications in Statistics: Theory and Methods*, 26:1481–96, 1997.

[7] E. McFowland, S. Speakman, and D. B. Neill. Fast generalized subset scan for anomalous pattern detection. *Journal of Machine Learning Research*, 14(1):1533–1561, 2013.

[8] E. McFowland, S. Speakman, and D. B. Neill. Fast generalized subset scan for anomalous pattern detection. *Journal of Machine Learning Research*, 14(1):1533–1561, 2013.

[9] D. B. Neill, A. W. Moore, M. Sabhnani, and K. Daniel. Detection of emerging space-time clusters. In R. Grossman, R. Bayardo, and K. P. Bennett, editors, *KDD*, pages 218–227. ACM, 2005.

[10] J. Qian, V. Saligrama, and Y. Chen. Connected sub-graph detection. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics, AISTATS 2014, Reykjavik, Iceland, April 22-25, 2014*, pages 796–804, 2014.

[11] P. Rozenshtein, A. Anagnostopoulos, A. Gionis, and N. Tatti. Event detection in activity networks. In *The 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '14, New York, NY, USA - August 24 - 27, 2014*, pages 1176–1185, 2014.

[12] J. Sharpnack, A. Krishnamurthy, and A. Singh. Near-optimal anomaly detection in graphs using lovasz extended scan statistic. In C. J. C. Burges, L. Bottou, Z. Ghahramani, and K. Q. Weinberger, editors, *NIPS*, pages 1959–1967, 2013.

[13] J. Sharpnack, A. Singh, and A. Rinaldo. Sparsistency of the edge lasso over graphs. In N. D. Lawrence and M. Girolami, editors, *AISTATS*, volume 22 of *JMLR Proceedings*, pages 1028–1036. JMLR.org, 2012.

[14] J. Sharpnack, A. Singh, and A. Rinaldo. Changepoint detection over graphs with the spectral scan statistic. In *AISTATS*, volume 31 of *JMLR Proceedings*, pages 545–553. JMLR.org, 2013.