T.R.
GEBZE TECHNICAL UNIVERSITY
FACULTY of ENGINEERING
DEPARTMENT of COMPUTER ENGINEERING

ANOMALY DETECTION PAPER RESEARCH

CSE 454 DATA MINING
ASSIGNMENT 2 REPORT

STUDENT
Şeyda Nur DEMİR
12 10 44 042

LECTURER
Burcu YILMAZ

TEACHING ASSISTANT
-

KOCAELİ, 2020

1.DESCRIPTION

1.1.Requirements :

1. You should read a conference paper related to anomaly detection.
   The paper should be from one of the international conferences.
   Papers from conferences held in Turkey will not be accepted.
   The conference should be one of the top conferences which is held after 2015.
2. You have to write a report in Turkish summarising the paper including, which should includes
   a. details about the conference,
   b. the method,
   c. results,
   d. a discussion about the results
   e. and the contribution of the paper to the literature.
3. You should upload the report and the paper to the moodle.

1.2.Deadline :

- ~~13.12.2020 23:55~~ (Delayed) 20.12.2020 23:55
- You must upload your assignment to moodle.

1.3.Demo :

- Before 15.01.2021 (Please ask an appointment from me to attend the demo)
- You must present your code in the demo section. If you do not attend the demo section, you will get zero from the assignment, even if you upload your assignment.

2.ANOMALY DETECTION PAPER RESEARCH

❖ What is anomaly detection?

Outlier detection (also known as anomaly detection) is the process of finding data objects with behaviors that are very different from expectation. Such objects are called outliers or anomalies. Outlier detection is important in many applications in addition to fraud detection such as medical care, public safety and security, industry damage detection, image processing, sensor/video network surveillance, and intrusion detection.

❖ What about my anomaly detection paper research?

KDD 2018

August 19-23, 2018, London, United Kingdom

https://www.kdd.org/kdd2018/

The annual KDD conference is the premier interdisciplinary conference bringing together researchers and practitioners from data science, data mining, knowledge discovery, large-scale data analytics, and big data. (ACM)

https://www.acm.org/

Watch on Youtube

https://www.youtube.com/watch?v=S8AhKd7h-hE

Read on Acm Doi

https://dl.acm.org/doi/10.1145/3219819.3220040

Title

SpotLight: Detecting Anomalies in Streaming Graphs

https://www.kdd.org/kdd2018/accepted-papers/view/spotlight-detecting-anomalies-in-streaming-graphs

Authors

Dhivya Eswaran, *deswaran@cs.cmu.edu*, Carnegie Mellon University, Pittsburgh, PA, USA

Christos Faloutsos, *christos@cs.cmu.edu*, Carnegie Mellon University, Pittsburgh, PA, USA

Sudipto Guha, *sudipto@amazon.com*, Amazon, New York City, NY, USA

Nina Mishra, *nmishra@amazon.com*, Amazon, Palo Alto, CA, USA

CCS Concepts

• Information systems → Data stream mining;

• Theory of computation→Graph algorithms analysis;

Keywords

• Anomaly detection

• Streaming graphs

• Graph sketching

Abstract

How do we spot interesting events from e-mail or transportation logs? How can we detect port scan or denial of service attacks from IP-IP communication data? In general, given a sequence of weighted, directed or bipartite graphs, each summarizing a snapshot of activity in a time window, how can we spot anomalous graphs containing the sudden appearance or disappearance of large dense subgraphs (e.g., near bicliques) in near real-time using sublinear memory? To this end, we propose a randomized sketching-based approach called SpotLight, which guarantees that an anomalous graph is mapped 'far' away from 'normal' instances in the sketch space with high probability for appropriate choice of parameters. Extensive experiments on real-world datasets show that SpotLight (a) improves accuracy by at least 8.4% compared to prior approaches, (b) is fast and can process millions of edges within a few minutes, (c) scales linearly with the number of edges and sketching dimensions and (d) leads to interesting discoveries in practice.

Contents

Acknowledgments

We thank Roger Barga, Charles Elkan, Nima Sharifi Mehr, Morteza Monemizadeh and Yonatan Naamad for insightful discussions.

Summary (in Turkish)

✓ Bu makalede, verilen bir weighted ve directed veya bipartite stream graphta, eğer varsa, gerçek zamana yakın, aniden görünen veya kaybolan, büyük dense directed subgraphları, sublinear memory kullanarak tespit etme problemi ele alınmıştır.

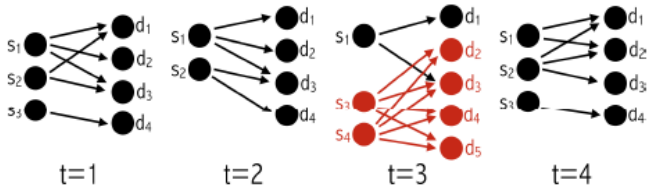

Figure 1: Sudden appearance of a dense subgraph at t=3.

✓ Bu problemin çözümü ile, e-maillerdeki veya transportation loglarındaki ilginç olaylar tespit edilebilir, IP-IP communication verilerinden, port scan veya denial of service saldırıları tespit edilebilir, ve bunun gibi gerçek zamanlı gerçekleşen anormal durumların tespiti sağlanabilir. Problemi önemli yapan en büyük faktör, anomalinin gerçek zamanda gerçekleşiyor olması ve gerçek zamanda algılanmasının gerektiği gerçeğidir. Gerçek zamanlı alınan verilerden anlık çizilen graphlarda, anomalinin tespit edilmesi, ağ güvenliği gibi konularda ciddi önem arz etmektedir.

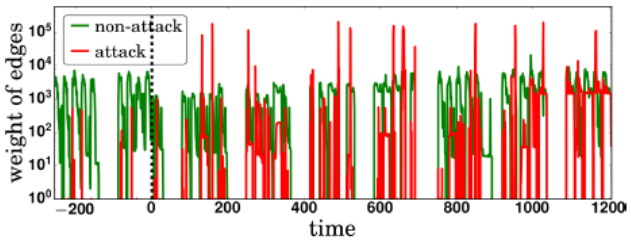

Figure 10: Understanding (un)detected anomalies in DARPA using the number of attack and non-attack edges over time.

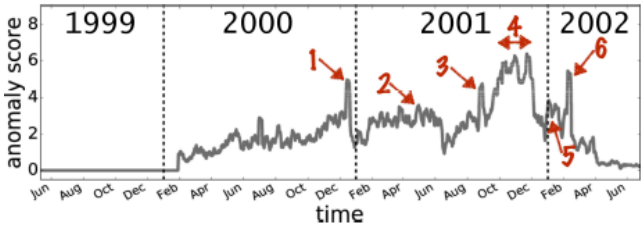

Figure 11: Anomaly detection on ENRON dataset

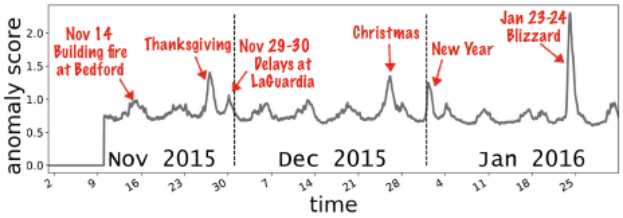

Figure 12: Anomaly detection on NYCTAXI dataset

✓ Makalede probleme önerilen çözüm ise, “Spotlight” yaklaşımıdır. Bu yaklaşım, çok yüksek olasılıklarla doğru parametreleri seçen, sketch space’de anormal graphları normal instance’lardan çok uzağa map’lemeyi garanti eden, randomized bir skecth-based yaklaşımdır.

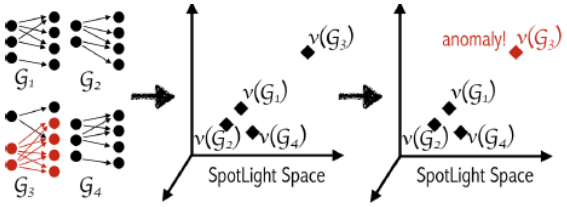

Figure 2: Overview of SpotLight

✓ Gerçek zamanlı gerçek verilerle yapılan kapsamlı deneyler göstermiştir ki Spotlight,
- Accuracy oranını birincil yaklaşımlara kıyasla en az %8.4 oranında geliştirir.
- Hızlıdır ve milyonlarca edge’i birkaç dakika içerisinde işler,
- Edge numaralarıyla lineer olarak ölçeklendirir ve alanları çizer,
- Pratikte ilginç keşiflere yol açar.

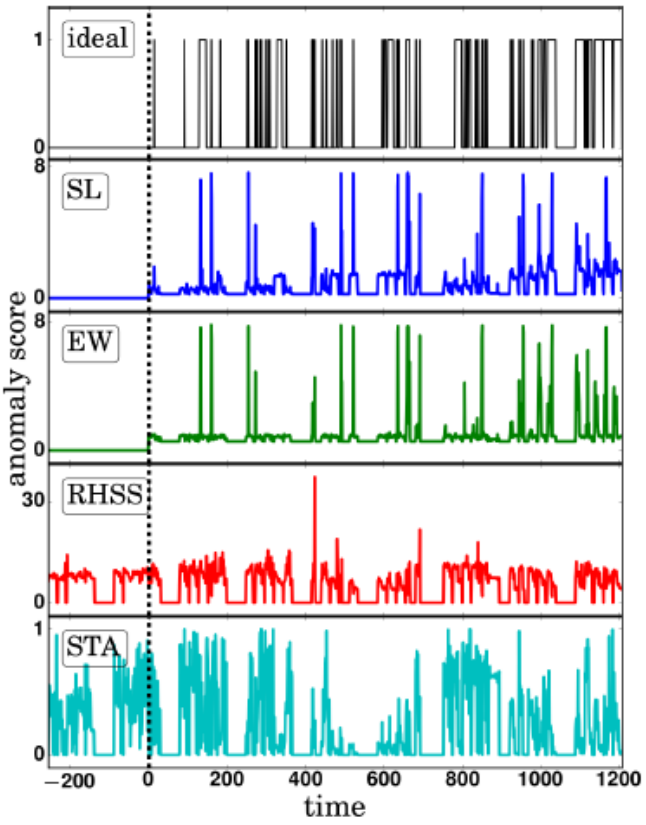

Figure 9: Anomaly detection on DARPA dataset. Spikes in the ‘ideal’ black curve indicate ground truth anomalies.

Spotlight Graph Stream Anomali Tespiti Algoritması

Algorithm 1 SPOTLIGHT graph stream anomaly detection

Input: a stream $\mathfrak{G}$ of weighted directed/bipartite graphs
Parameters: sketch dimensionality K, source sampling probability p, destination sampling probability q
Output: a stream of anomaly scores

1: **procedure** SPOTLIGHT($\mathfrak{G}, K, p, q$)
2: INITIALIZE(K, p, q)
3: **for** graph $\mathcal{G} \in \mathfrak{G}$ **do**
4: $\mathbf{v} \leftarrow$ SKETCH($\mathcal{G}$)
5: **yield** ANOMALYSCORE($\mathbf{v}$)
6: **procedure** INITIALIZE(K, p, q)
7: **for** $k = 1, \ldots, K$ **do**
8: Pick source hash $h_k : \mathcal{S} \rightarrow \{1, \ldots, \lfloor 1/p \rfloor\}$ and destination hash $h'_k : \mathcal{D} \rightarrow \{1, \ldots, \lfloor 1/q \rfloor\}$ independently at random.
9: **procedure** SKETCH($\mathcal{G}$)
10: $\mathbf{v} \leftarrow \mathbf{0}_K$
11: **for** edge $e = (s, d, w)$ in graph $\mathcal{G}$ **do**
12: **for** $k = 1, \ldots, K$ **do**
13: **if** $h_k(s) == 1$ and $h'_k(d) == 1$ **then**
14: $v_k \leftarrow v_k + w$
15: **return** $\mathbf{v}$

- ✓ Bu metoda göre, her G graphı için K-boyutlu bir SpotLight taslağı v(G) çıkarılır, öyle ki, büyük dense subgraphların ani görünümün veya kayboluşunu içeren graphlar, sketch space'deki (satır 4) normal grafiklerden çok uzaktır.

- ✓ Ve, anormal graphlar olarak anormal çizimler veren grafikleri tespit etmek için sketch space'teki distance gap'i kullanır (satır 5).

✓ Gerçek zamanlı gerçek veri testleri ile gerçekleştirilen testler, Spotlight'ın çalışırlığını ve geliştirmelerini doğrulamıştır. Spotlight, diğer yaklaşımlarla kıyaslandığında, en iyi performansı göstermiştir.

Method	precision@ 100	200	300	400	recall@ 100	200	300	400
Ideal	*1.0*	*1.0*	*0.96*	*0.72*	*0.35*	*0.69*	*1.0*	*1.0*
SL	**0.96**	**0.79**	**0.64**	**0.57**	**0.34**	**0.55**	**0.67**	**0.80**
EW	0.86	0.54	0.47	0.46	0.30	0.38	0.49	0.65
RHSS	0.31	0.28	0.32	0.36	0.11	0.19	0.33	0.50
STA	0.23	0.16	0.19	0.24	0.08	0.11	0.20	0.34

Table 2: SPOTLIGHT (SL) achieves better precision and recall than baselines (EW, RHSS, STA). Bold indicates the highest value in each column (excluding ideal). Underline shows significant differences (p-value ≤ 0.01) w.r.t. baselines according to a two-sided micro-sign test [30].

✓ Spotlight, diğer yaklaşımlara kıyasla en optimize sonucu vermiştir.

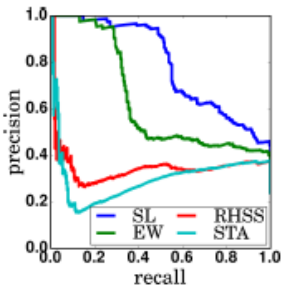

Figure 5: SL outperforms baselines in terms of precision and recall.

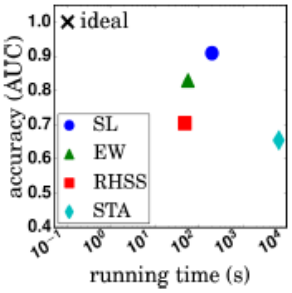

Figure 6: Accuracy-running time trade-off offered by SL.

✓ Spotlight'ın test sonuçları, yaklaşımın amacına ulaştığının göstermektedir.

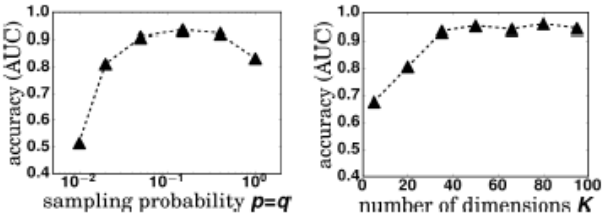

Figure 7: Variation of accuracy with (a) $p = q$ for $K = 10$ and (b) with K for $p = q = 0.1$.

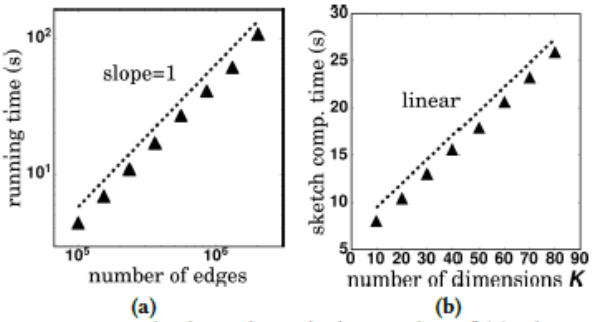

Figure 8: SL scales linearly with the number of (a) edges in the stream and (b) sketch dimensions K.

- ✓ Spotlight algoritmasının literatüre kazandırdığı somut kazanımları

  Algoritması, ele aldığı probleme karşı geliştirdiği radomized sketch-based yaklaşımı,

  Verdiği güvence, çok yüksek olasılıklarla doğru parametreleri seçmesi, sketch space'de anormal graphları normal instance'lardan çok uzağa map'lemeyi garanti etmesi,

  Efektif olması, gerçek zamanlı gerçek verilerle yapılan kapsamlı deney sonuçlarına göre, hızlı olması, zaman kazandırması, ölçülebilir olmasıdır.

- ✓ Spotlight, her ne kadar en hassas sonuçları verse de, her zaman daha iyisi vardır. Uyarlanabilir veriye dayalı eskizleri analiz etmesi zor olsa da, farklı yaklaşımlar daha iyi sonuç verebilir. Yine de şu an için basit olması, ölçülebilir olması, kodlama kolaylığı ve performansı açısından Spotlight'ı önerilir.

- ✓ Son olarak, anomalinin yörüngesi, onu anlamada ve tahmin etmede, hayati önem taşır.

END OF THE REPORT

UPDATED : 20.12.2020 23:10

STUDENT
Şeyda Nur DEMİR
12 10 44 042

LECTURER
Burcu YILMAZ

TEACHING ASSISTANT
-

KOCAELİ, 2020