SEDANSPOT: Detecting Anomalies in Edge Streams
Supplementary

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

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

***Abstract*—This document supplements [1] by giving proofs of theoretical results and details on additional experiments.**

I. PROOFS OF THEORETICAL ANALYSIS IN SEC. VI

Proof of Lem. 1: Follows from Proposition 3 of [2]. ■

Proof of Thm. 1: A sampled edge e can belong to $\mathcal{H}_k$ in $|\mathcal{H}_k|$ mutually exclusive ways. Hence, $\mathbf{Pr}\left[e \in \mathcal{H}_k \mid e \in \mathcal{S}\right] = \sum_{e' \in \mathcal{H}_k} \mathbf{Pr}\left[e = e' \mid e \in \mathcal{S}\right]$, which can be simplified using Bayes' rule to $\sum_{e' \in \mathcal{H}_k} \mathbf{Pr}\left[e = e' \wedge e \in \mathcal{S}\right] / \mathbf{Pr}\left[e \in \mathcal{S}\right] \propto \sum_{e' \in \mathcal{H}_k} \mathbf{Pr}\left[e' \in \mathcal{S}\right]$ since $\mathbf{Pr}\left[e \in \mathcal{S}\right]$ is independent of k.

Suppose $\gamma_1 < \gamma_2 \ldots < \gamma_{z_k} = \tau_k$ are the anchored time ticks during interval $I_k := (\tau_{k-1}, \tau_k]$. Let $\gamma_0 = \tau_{k-1}$ and $\mathcal{E}_{\gamma_i}$ be the set of edges occuring at γ_i. Then, $\sum_{e' \in \mathcal{E}_{\gamma_i}} \mathbf{Pr}\left[e' \in \mathcal{S}\right] \propto \sum_{e' \in \mathcal{E}_{\gamma_i}} 1/r(e') = \gamma_i - \gamma_{i-1}$. Thus, $\mathbf{Pr}\left[e \in \mathcal{H}_k \mid e \in \mathcal{S}\right] \propto \sum_{e' \in \mathcal{H}_k} \mathbf{Pr}\left[e' \in \mathcal{S}\right] = \sum_{i=1}^{z_k} \gamma_i - \gamma_{i-1} = \tau_k - \tau_{k-1} = \ell_k$. ■

Proof of Lem. 2: $\mathcal{O}\left(N/\alpha\right)$ comes from the N local random walks, each of expected length $\mathbb{E}\left[\text{Geometric}(\alpha)\right] = 1/\alpha$ (lines 16-22 of Alg. 3). Each step of random walk takes $\mathcal{O}\left(1\right)$ due to constant-time neighbor sampling via LAT. ■

Proof of Lem. 3: The $\mathcal{O}\left(\log|\mathcal{V}|\right)$ is a lower bound on the memory requirement, since each edge (including vertex IDs) needs to be read. Thus, $\mathcal{O}\left(S \log|\mathcal{V}|\right)$ space is needed to store LAT data structures over the sample of S edges. ■

Proof of Lem. 4: For $i > S$, the probability of sampling the i^{th} edge e_i in the stream is $p_i = r(e_i)^{-1} / \sum_{j=1}^{i} r(e_j)^{-1} \leq r_{\max}/(i \cdot r_{\min})$. The expected number of updates is $S + \sum_{S+1}^{L} p_i \leq S + r_{\max}/r_{\min} \sum_{S+1}^{L} 1/i$, each update costing $\mathcal{O}\left(\log S\right)$ for sampling and $\mathcal{O}\left(d_{avg}\right)$ for scoring in expectation (assuming no correlation between edge rate and the degrees of incident vertices). Amortization completes the proof. ■

II. ADDITIONAL EXPERIMENTS

Here, we describe our datasets and then answer the following from the main paper: *Q3) Discoveries:* Does SEDANSPOT lead to interesting discoveries in practice? *Q4) Parameters:* How do accuracy and running time depend on S, N and α?

A. Dataset description

DARPA [3] dataset consists of network traffic from 9484 source IPs to 23398 destination IPs over $87.7K$ minutes. There are $\sim 4.5M$ directed <*srcIP, dstIP, 1, time*> edges in total, of which 60% are manually annotated as anomalous. They correspond to 89 network attacks – such as denial of service or port scan – injected by domain experts. Despite this high proportion, the attacks themselves occurred infrequently (but as bursts of activity) and originated from a mix of IP addresses which either were solely dedicated to attacks, or more challengingly, attempted camouflage by participating in normal traffic. This makes DARPA dataset the perfect testbed for SEDANSPOT, which aims to detect precisely such anomalous bursts occurring in sparse regions of the graph.

ENRON [4] dataset consists of e-mail communications among the 151 employees of Enron company from May 1999 to April 2002, a period of three years surrounding the famous Enron scandal. There are $\sim 50K$ directed <*sender, receiver, 1, date*> edges. Since ground truth is not directly available, we verify anomalies by correlating their time stamps with real-world events. We expect more edges to be flagged anomalous during periods of large internal (e.g., new CEO) or external (e.g., updates on lawsuit) changes which create (or result from) excitement or turbulence among the employees.

DBLP [5] is the collaboration network of authors of papers from DBLP computer science bibliography. Each undirected edge <*auth1, auth2, 1, pub_year*> between two authors represents a joint publication. For simplicity, we only consider papers published in 1991-2010 and retain nodes (authors) who have at least 50 edges, filtering out the remaining nodes (and corresponding edges). This resulted in a graph containing around $55.5K$ authors and $3.7M$ coauthorships. We expect anomalous edges to represent unlikely collaborations, e.g., authors from unrelated fields or different geographical regions. We verify anomalies using the public profiles of the authors.

Q3) Discoveries (on ENRON *and* DBLP *datasets)*

ENRON: Fig. 1 depicts the anomaly detection results on ENRON dataset using N=50 walks, α=0.15 restart probability and S=$2K$ edges in sample. The inset, plotting the distribution of anomaly scores, showcases a large separation in scores for anomalies and non-anomalies, which is desirable. Accordingly, we use the threshold marked in red (=0.76) to flag anomalies[1]. The rest of the figure plots the temporal distribution of flagged edges (aggregated weekly), which we verify by correlating with the publicly-available ENRON time line[2].

The red arrows in Fig. 1 mark the top five non-contiguous periods of time having the highest number of anomalous edges. As expected, these periods coincide well with notable events

[1] In practice, one can use the median $\hat{\mu}$ and inter-quantile range $\hat{\sigma}$ of past scores to flag anomalies, *in an online manner*, when the score exceeds $\hat{\mu} + 3\hat{\sigma}$.

[2] http://www.agsm.edu.au/bobm/teaching/BE/Enron/timeline.html

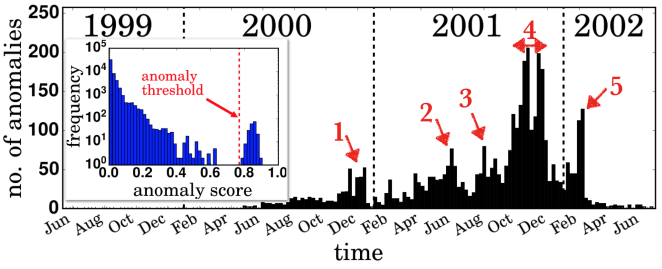


Fig. 1. Anomaly detection in ENRON dataset

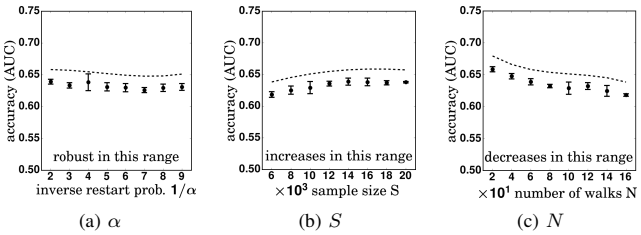
surrounding the ENRON scandal, creating a flood of unusual e-mails from (even low-level) employees: (1) Dec 2000: Skilling announced as CEO. (2) Jun 2001: The California energy crisis ends. (3) Aug 2001: Skilling announces resignation. Lay named CEO. (4) Oct-Nov 2001: Fastow ousted. A formal investigation against Enron is launched. Stocks crumble. Enron files for bankruptcy. (5) Jan-Feb 2002: Cooper takes over as CEO after Lay resigns. Fastow, Kopper, Skilling and Watson (whistle-blower) testify before Congress.

DBLP: We run SEDANSPOT on DBLP with $S{=}200K$, $\alpha{=}0.15$ and $N{=}1000$. As we show with the help of anecdotal evidence, the top anomalous edges indeed represent unexpected or unlikely collaborations – *(i) Alex Galis, Robert Szabo (2004):* This is due a joint invited paper at an IEEE MATA 2004 workshop, which marked the beginning of an unexpected collaboration between authors of different countries, namely, Galis from Univ. College London (UK) and Szabo from Budapest Univ. of Technology and Economics (Hungary). *(ii) Nikol Rummel, Nikolaos Avouris (2007):* This is the result of an interdisciplinary paper about Computer Supported Collaborative Learning, requiring collaboration between authors belonging to different fields (Psychology, ECE). *(iii) Ryan Thibodeau, Mark Carrington (2010):* This is the product of a rare massive collaboration effort among 44 authors across seven institutions, including Thibodeau from Univ. of Georgia, USA and Carrington from Univ. of Cambridge, UK. This marks their only joint publication.

Q4) Accuracy and running time w.r.t. parameters

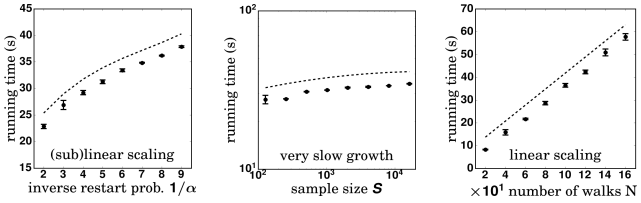
Fig. 2 and Fig. 3 show how accuracy and running time (to process $0.5M$ edges) vary with the number of walks N, restart probability α and sample size S. By default, we use $S{=}10K$, $N{=}100$ and $\alpha{=}0.15$. All values are averaged over five runs and error bars indicate standard deviations.

Accuracy w.r.t. α: Fig. 2a shows that the accuracy is robust (~ 0.635) to the restart probability $\alpha \in [0.5, 0.11]$. This is consistent with the trend observed for page rank (closely related to random walks), where α has little effect on the top ranking webpages based on their page ranks [6]. **W.r.t. S:** Fig. 2b shows that the accuracy exhibits a 'diminishing return' behavior as sample size S is increased in $[6K, 20K]$. As SEDANSPOT stores more edges, it better models normal behavior and thus accuracy increases. But the marginal increase itself decreases: once the normal behavior is captured



(a) α (b) S (c) N

Fig. 2. Accuracy w.r.t. parameters



(a) α (b) S (c) N

Fig. 3. Running time w.r.t. parameters

sufficiently well in the sample, subsequent increase in S leads to little improvement. **W.r.t. N:** Fig. 2c, somewhat surprisingly, shows that increasing the number of walks N did not necessarily lead to higher accuracy. In fact, in further experiments, we found that the accuracy peaks around $N{=}10$ and then gradually starts decreasing. A similar pattern has been observed for tasks like link prediction, where estimates of RWR relevance scores based on a few walks often outperform their steady state values [7].

Running time w.r.t. α: Fig. 3b plots the running time of SEDANSPOT against $1/\alpha$, where the restart probability $\alpha \in [1/2, 1/9]$. The curve begins to flatten out for high values of $1/\alpha$, suggesting that the running time scales sublinearly with $1/\alpha$. This is natural given the finite sample of edges that SEDANSPOT maintains: even though the expected length of walks increases linearly with $1/\alpha$, many of these walks terminate early, resulting in a sublinear dependence. **W.r.t. S:** Fig. 3c shows that the running time grows very slowly with sample size S in $[128, 16384]$. The running time is small for low S due to premature termination of local random walks in a small sample of edges (many vertices did not have any out-edges). **W.r.t. N:** Fig. 3a shows that SEDANSPOT scales linearly with N as the points align well on a straight line.

REFERENCES

[1] D. Eswaran and C. Faloutsos, "Sedanspot: Detecting anomalies in edge streams," in *ICDM*. IEEE, 2018.

[2] P. S. Efraimidis and P. G. Spirakis, "Weighted random sampling with a reservoir," *Inf. Process. Lett.*, vol. 97, no. 5, pp. 181–185, 2006.

[3] R. Lippmann, R. K. Cunningham, D. J. Fried, I. Graf, K. R. Kendall, S. E. Webster, and M. A. Zissman, "Results of the DARPA 1998 offline intrusion detection evaluation," in *Recent Advances in Intrusion Detection*, 1999.

[4] J. Shetty and J. Adibi, "The enron email dataset database schema and brief statistical report," *Information sciences institute technical report, University of Southern California*, vol. 4, no. 1, pp. 120–128, 2004.

[5] "Dblp network dataset," http://konect.uni-koblenz.de/networks/dblp_coauthor, 2014.

[6] A. N. Langville and C. D. Meyer, "Deeper inside pagerank," *Internet Mathematics*, vol. 1, no. 3, pp. 335–380, 2004.

[7] W. Liu and L. Lü, "Link prediction based on local random walk," *EPL (Europhysics Letters)*, vol. 89, no. 5, p. 58007, 2010.