Beyond Outlier Detection: LookOut for Pictorial Explanation

Nikhil Gupta[1(✉)], Dhivya Eswaran[2], Neil Shah[3], Leman Akoglu[2], and Christos Faloutsos[2]

[1] IIT Delhi, New Delhi, India
Nikhil.Gupta.cs514@cse.iitd.ac.in
[2] CMU, Pittsburgh, USA
{deswaran,lakoglu,christos}@cs.cmu.edu
[3] Snap Inc., Santa Monica, USA
nshah@snap.com

Abstract. Why is a given point in a dataset marked as an outlier by an off-the-shelf detection algorithm? Which feature(s) explain it the best? What is the best way to convince a human analyst that the point is indeed an outlier? We provide succinct, interpretable, and simple pictorial explanations of outlying behavior in multi-dimensional real-valued datasets while respecting the limited attention of human analysts. Specifically, we propose to output a few focus-plots, i.e., pairwise feature plots, from a few, carefully chosen feature sub-spaces. The proposed LookOut makes four contributions: (a) **problem formulation:** we introduce an "analyst-centered" problem formulation for explaining outliers via focus-plots, (b) **explanation algorithm:** we propose a plot-selection objective and the LookOut algorithm to approximate it with optimality guarantees, (c) **generality:** our explanation algorithm is *both* domain- and detector-agnostic, and (d) **scalability:** LookOut scales linearly with the size of input outliers to explain and the explanation budget. Our experiments show that LookOut performs near-ideally in terms of maximizing explanation objective on several real datasets, while producing visually interpretable and intuitive results in explaining groundtruth outliers. Code related to this paper is available at: https://github.com/NikhilGupta1997/Lookout.

Keywords: Outlier detection · Pictorial explanation · Interpretability

1 Introduction

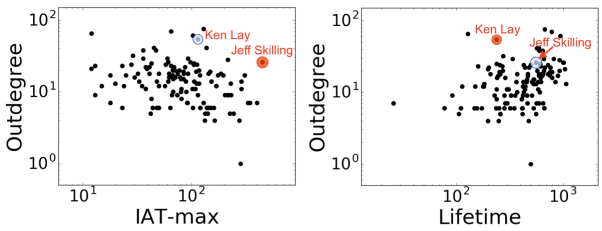
Given a multi-dimensional dataset of real-valued features, e.g., sensor measurements, and a list of outliers (identified by an off-the-shelf "black-box" detector or any other external mechanism), how can we *explain* the outliers to a human analyst in a succinct, effective, and interpretable fashion?

Outlier detection is a widely studied problem. Numerous detectors exist for point data [1,5,18], time series [12], as well as graphs [2,3]. However, the literature on outlier explanation is surprisingly sparse. Given that the outcomes

M. Berlingerio et al. (Eds.): ECML PKDD 2018, LNAI 11051, pp. 122–138, 2019.
https://doi.org/10.1007/978-3-030-10925-7_8

Name	Score
Skilling, Jeff	0.893
Lay, Kenneth	0.761
Fastow, Andrew	0.442
Mark, Rebecca	0.429
Smith, John	0.331
Cooper, Stephen	0.308
Tomson, Mary	0.232
...	...

(a) traditional



(b) proposed

Fig. 1. Compared to traditional ranked list output (a: *wordy, lengthy, no explanation*), LookOut produces simple, interpretable explanations (b: *visual, succinct, interpretable*) (Color figure online)

(alerts) of a detector often go through a "vetting" procedure by human analysts, it is beneficial to provide explanations for such alerts which can empower analysts in sense-making and reduce their efforts in troubleshooting. Moreover, such explanations should justify the outliers succinctly in order to save analyst time.

Our work sets out to address the above outlier explanation problem. Consider the following **example situation:** Given performance metrics from hundreds of machines within a large company, an analyst could face two relevant scenarios.

- *Detected outliers:* For monitoring, s/he could use any "black-box" outlier detector to spot machines with suspicious values of metric(s). Here, we are oblivious to the specific detector, knowing only that it flags outliers, but does not produce any interpretable explanation.
- *Dictated outliers:* Alternatively, outlying machines may get reported to the analyst externally (e.g., they crash or get compromised).

In both scenarios, the analyst would be interested in understanding *in what ways* the pre-identified outlying machines (detected or dictated) differ from the rest.

In this work, we propose a new approach called LookOut, for explaining a given set of outliers, and apply it to various, relational and non-relational, settings. At its heart, LookOut provides interpretable *pictorial* explanations through simple, easy-to-grasp focus-plots (Definition 1), which "incriminate" the given outliers the most. We summarize our contributions as follows.

- **Outlier Explanation Problem Formulation:** We introduce a new formulation that explains outliers through "focus-plots". In a nutshell, given the list of outliers from a dataset with real-valued features, we aim to find a few 2D plots on which the total outlier "blame" is maximized. Our emphasis is on two key aspects: (a) *interpretability:* our plots visually incriminate the outliers, and (b) *succinctness:* we show only a few plots to respect the analyst's attention; the analysts can then quickly interpret the plots, spot the outliers, and verify their abnormality given the discovered feature pairs.
- **Succinct Quantifiable Explanation Algorithm LookOut:** We propose the LookOut algorithm to solve our explanation problem. Specifically, we

develop a plot selection objective, which quantifies the 'goodness' of an explanation and lends itself to monotone submodular function optimization, which we solve efficiently with optimality guarantees. Figure 1 illustrates LookOut's performance on the Enron communications network, where it discovers two focus-plots which maximally incriminate the given outlying nodes: Enron founder "Ken Lay" and CEO "Jeff Skilling." Note that the outliers stand out visually from the normal nodes.

- **Generality:** LookOut is general in two respects: it is (a) *domain-agnostic*, meaning it is suitable for datasets from various domains, and (2) *detector-agnostic*, meaning it can be employed to explain outliers produced by any detector or identified through any other mechanism (e.g., crash reports).
- **Scalability:** We show that LookOut requires time linear on (i) the number of plots to choose explanations from, (ii) the number of outliers to explain and (iii) the user-specified budget for explanations (see Lemma 7 and Fig. 5).

We experiment with several real-world datasets from diverse domains including e-mail communications and astronomy, which demonstrate the effectiveness, interpretability, succinctness and generality of our approach.

Reproducibility: Our datasets are publicly available (See Sect. 5.1) and LookOut is open-sourced at https://github.com/NikhilGupta1997/Lookout.

2 Related Work

While there is considerable prior work on outlier detection [3,6,12], literature on outlier description is comparably sparse. Several works aim to find an optimal feature subspace which distinguishes outliers from normal points. [14] aims to find a subspace which maximizes differences in outlier score distributions of all points across subspaces. [17] instead takes a constraint programming approach which aims to maximize differences between neighborhood densities of known outliers and normal points. An associated problem focuses on finding minimal, or optimal feature subspaces for each outlier. [15] aims to give "intensional knowledge" for each outlier by finding minimal subspaces in which the outliers deviate sufficiently from normal points using pruning rules. [7,8] use spectral embeddings to discover subspaces which promote high outlier scores, while aiming to preserve

Properties vs. Methods	Knorr et al. [15]	Dang et al.[7]	Angiulli et al. [4]	Micenkova et al. [19]	Kopp et al. [16]	Keller et al. [14]	LookOut
Quantifiable explanations		✔	✔	✔		✔	✔
Budget-conscious						✔	✔
Visually interpretable							✔
Scalable	✔				✔	✔	✔

Fig. 2. Comparison with other outlier description approaches, in terms of four desirable properties

distances of normal points. [19] instead employs sparse classification of an inlier class against a synthetically-created outlier class for each outlier in order to discover small feature spaces which discern it. [16] proposes combining decision rules produced by an ensemble of short decision trees to explain outliers. [4] augments the per-outlier problem to include outlier groups by searching for single features which differentiate many outliers.

All in all, none of these works meet several key desiderata for outlier description: (a) quantifiable explanation quality, (b) budget-consciousness towards analysts, (c) visual interpretability, and (d) a scalable descriptor, which is subquadratic on the number of nodes and at worst polynomial on (low) dimensionality. Figure 2 shows that unlike existing approaches, our LookOut approach is designed to give quantifiable explanations which aim to maximize *incrimination*, respect human attention-budget and visual interpretability constraints, and scale linearly on the number of outliers.

3 Preliminaries and Problem Statement

3.1 Notation

Let $\mathcal{V}$ be the set of input data points, where each point $v \in \mathcal{V}$ originates from $\mathbb{R}^d$ and $n = |\mathcal{V}|$ is the total number of points. Here, $d = |\mathcal{F}|$ is the dimensionality of the dataset and $\mathcal{F} = \{f_1, f_2, \ldots, f_d\}$ is the set of real-valued features (either directly given, or extracted, e.g., from a relational dataset). The set of outlying points given as input is denoted by $\mathcal{A} \subseteq \mathcal{V}$, $|\mathcal{A}| = k$. Typically, $k \ll n$. Table 1 summarizes the frequently used notation.

3.2 Intuition and Proposed Problem

The explanations we seek to generate should be simple, interpretable, and easy to illustrate to humans who will ultimately leverage the explanations. To this end, we decide to use *focus-plots* (Definition 1) for outlier justification, due to their visual appeal and interpretability. A formal definition is given below.

Definition 1 (Focus-plot). *Given a dataset of points $\mathcal{V}$, a pair of features $f_x, f_y \in \mathcal{F}$ (where $\mathcal{F}$ is the set of real-valued features) and an input set of outliers $\mathcal{A}$, focus-plot $p \in \mathcal{P}$ is a 2-d scatter plot of all points, with f_x on x-axis, f_y on y-axis, 'drawing attention' to the set of "maxplained" (maximally explained) outliers $\mathcal{A}_p \subseteq \mathcal{A}$ best explained by this feature pair.*

Intuitively, our pictorial outlier explanation is a set of *focus-plots*, each of which "blames" or "explains away" a subset of the input outliers, whose outlierness is best showcased by the corresponding pair of features. That is, we consider $\binom{d}{2} = \frac{d(d-1)}{2}$ 2-d spaces by generating all pairwise feature combinations. Within each 2-d space, we then score the points in $\mathcal{A}$ by their outlierness (Sect. 4.1).

Let us denote the set of all $\binom{d}{2}$ focus-plots by $\mathcal{P}$. Even for small values of d, showing all the focus-plots would be too overwhelming for the analyst. Moreover, some outliers could redundantly show up in multiple plots. Ideally, we

would identify only a few focus-plots, which could "blame" or "explain away" the outliers to the largest possible extent. In other words, our goal would be to output a small subset $\mathcal{S}$ of $\mathcal{P}$, on which points in $\mathcal{A}$ receive high outlier scores (Sect. 4.2). Given this intuition, we formulate our problem below.

Table 1. Symbols and definitions

Symbol	Definition
$\mathcal{V}$	Set of data points, $\lvert\mathcal{V}\rvert = n$
$\mathcal{A}$	Input set of outliers, $\lvert\mathcal{A}\rvert = k$
$\mathcal{F}$	Set of features, $\lvert\mathcal{F}\rvert = d$
$\mathcal{P}$	Set of focus-plots, $\lvert\mathcal{P}\rvert = d(d-1)/2 = l$
$s_{i,j}$	Outlier score of $a_i \in \mathcal{A}$ in plot $p_j \in \mathcal{P}$
$\mathcal{S}$	Subset of selected focus-plots
$f(\mathcal{S})$	Explanation objective function
$\Delta_f(p \mid \mathcal{S})$	Marginal gain of plot p w.r.t $\mathcal{S}$
b	Budget, i.e., maximum cardinality of $\mathcal{S}$

Problem 2 (Outlier Explanation).

- **Given** (a) a dataset on points $\mathcal{V}$ consisting of real-valued features $\mathcal{F}$, (b) a list of outliers $\mathcal{A} \subseteq \mathcal{V}$, either (1) detected by an off-the-shelf detector or (2) dictated by external information, and (c) a fixed budget of b focus-plots,
- **find** the *best* such focus-plots $\mathcal{S} \subseteq \mathcal{P}$, $\lvert\mathcal{S}\rvert = b$, so as to *to maximize* the *total maximum outlier score of outliers* that we can "blame" through the b plots.

4 Proposed Algorithm LookOut

In this section, we detail our approach for scoring the input outliers by focus-plots, the overall complexity analysis of LookOut, and conclude with discussion.

4.1 Scoring by Feature Pairs

Given all the points $\mathcal{V}$, with marked outliers $\mathcal{A} \subseteq \mathcal{V}$, and their given (or extracted) features $\mathcal{F} \in \mathbb{R}^d$, our first step is to quantify how much "blame" we can attribute to each input outlier in $\mathbb{R}^2$. As previously mentioned, 2-d spaces are easy to illustrate visually with focus-plots. Moreover, outliers in 2-d are easy to interpret: e.g., "point a has too many/too few $y =$ dollars for its $x =$ number of accounts". Given a focus-plot, an analyst can easily discern the outliers visually and come up with such explanations without any further supervision.

We construct 2-d spaces (f_x, f_y) by pairing the features $\forall x, y = \{1, \ldots, d\}, x \neq y$ (order does not matter). Each focus-plot $p_j \in \mathcal{P}$ corresponds to such a pair of features, $j = \{1, \ldots, \binom{d}{2}\}$. For scoring, we consider two different scenarios, depending on how the input outliers were obtained.

If the outliers are *detected* by some "black-box" detector available to the analyst, we can employ the same detector on all the nodes (this time in 2-d) and thus obtain the scores for the nodes in $\mathcal{A}$.

If the outliers are *dictated*, i.e. reported externally, then the analyst could use any off-the-shelf detector, such as LOF [5], DB-outlier [15], etc. In this work, we use the Isolation Forest (iForest) detector [18] for two main reasons: (a) it boasts *constant* training time and space complexity (i.e., independent of n) due to its sampling strategy, and (b) it has been shown empirically to outperform alternatives [9] and is thus state-of-the-art. However, note that none of these existing detectors has the ability to *explain* the outliers, especially iForest, as it is an ensemble approach.

By the end of the scoring process, each outlier receives $|\mathcal{P}| = l = \binom{d}{2}$ scores.

4.2 Plot Selection Objective

While scoring in small, 2-d spaces is easy and can be trivially parallelized, presenting all such focus-plots to the analyst would not be productive given their limited attention budget. As such, our next step is to carefully select a short list of plots that best blame all the outliers collectively, where the plot budget can be specified by the user.

Objective Function. While selecting plots, we aim to incorporate the following criteria:

- **incrimination power;** such that the outliers are scored as highly as possible,
- **high expressiveness;** where each plot incriminates multiple outliers, so that the explanation is *sublinear* in the number of outliers, and
- **low redundancy;** such that the plots do not explain similar sets of outliers.

We next introduce our objective criterion which satisfies the above requirements.

At this step of the process, we can conceptually think of a complete, weighted bipartite graph between the k input outliers $\mathcal{A} = \{a_1, \ldots, a_k\}$ and l focus-plots $\mathcal{P} = \{p_1, \ldots, p_l\}$, in which edge weight $s_{i,j}$ depicts the outlier score that a_i received from p_j, as illustrated in Fig. 3.

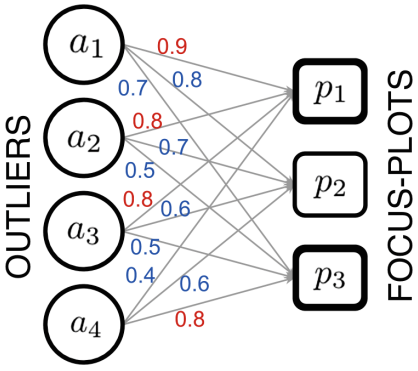


Fig. 3. LookOut with $k = 4$ outliers, $l = 3$ focus-plots, and budget $b = 2$: p_1 is picked first due to maximum *total incrimination* (sum of edge weights = 2.9); next p_3 is chosen over p_2, due to its higher marginal gain (0.4 vs 0.2) (Color figure online)

Algorithm 1. LookOut

Data: dataset of points $\mathcal{V}$, outliers $\mathcal{A}$, set of all focus-plots $\mathcal{P}$, budget b
Result: pictorial outlier explanation $\mathcal{S}$, which is a set of focus-plots

```
for p_j ∈ P do
    D_j ← iForest constructed using V and the two features in plot p_j;
    for a_i ∈ A do
        s_{i,j} ← anomaly score given by detector D_j to point a_i;
    end
end
initialize S ← ∅;
while |S| < b do
    recompute marginal gain Δ_f(p | S) ∀ p ∈ P \ S ;      // using Eq. 4
    p* ← arg max_{p∈P\S} Δ_f(p | S) ;
    S ← S ∪ {p*};
end
return S ;
```

We formulate our objective to *maximize* the total *maximum outlier score* of each outlier amongst the selected plots:

$$\underset{\mathcal{S}\subseteq\mathcal{P},|\mathcal{S}|=b}{\text{maximize}} \quad f(\mathcal{S}) = \sum_{a_i\in\mathcal{A}} \max_{p_j\in\mathcal{S}} s_{i,j} \tag{1}$$

Here, our objective function, $f(\mathcal{S})$, can be considered the *total incrimination* score given by subset $\mathcal{S}$. Since we are limited with a budget of plots, we aim to select those which explain multiple outliers to the best extent. Note that each outlier receives their maximum score from exactly one of the plots among the selected set (excluding ties), which effectively *partitions* the explanations and avoids redundancy. In the example from Fig. 3, focus-plots p_1 and p_3 "explain away" outliers $\{a_1, a_2, a_3\}$ and $\{a_4\}$ respectively, where the maximum score that each outlier receives is highlighted in red font.

Concretely, we denote by $\mathcal{A}_p$ the set of *maxplained* (maximally explained) outliers by focus-plot p, i.e., outliers that receive their highest score from p, i.e. $\mathcal{A}_p = \{a_i | p = \arg\max_{p_j\in\mathcal{S}} s_{i,j}\}$, where we break ties at random. Note that $\mathcal{A}_p \cap \mathcal{A}_{p'} = \emptyset$, $\forall\ p, p' \in \mathcal{P}$. In depicting a plot p to the analyst, we mark the set of maxplained outliers $\mathcal{A}_p$ in red and the rest in $\mathcal{A}\backslash\mathcal{A}_p$ in blue – see Fig. 1.

4.3 Approximation Algorithm LookOut

Having defined our plot selection objective, we need to devise a subset selection algorithm to optimize Eq. (1), for a budget b. Notice that the optimal subset selection is a combinatorial task which we can show to be **NP-hard**.

Lemma 3. *The focus-plot selection problem in Eq. (1) is NP-hard.*

Proof. We sketch the proof by a reduction from the Maximum Coverage (MaxCover) problem, which is known to be NP-hard [10]. An instance of MaxCover involves an integer k and a collection of sets $\{S_1, \ldots, S_l\}$ each containing a list of elements, where the goal is to find k sets such that the total number of covered elements is maximized. The MaxCover problem instance maps to an instance of our problem, where each set S_j corresponds to a focus-plot p_j, each element e_i maps to an outlier a_i, and the elements (outliers) inside each set has the same unit score ($s_{i,j} = 1$ for $e_i \in S_j$) while the others outside the set has score zero ($s_{i,j} = 0$ for $e_i \notin S_j$) on the corresponding focus-plot. Since MaxCover is equivalent to a special case of our problem, we conclude that Eq. (1) is at least as hard as MaxCover. □

Therefore, our aim is to find an approximation algorithm to optimize Eq. (1).

Properties of Our Objective. Fortunately, our objective $f(\cdot)$ exhibits three key properties that enable us to use a greedy algorithm with an approximation guarantee. Specifically, our objective $f : 2^{|\mathcal{P}|} \to \mathbb{R}^+ \cup \{0\}$ is (i) *non-negative*, since the outlier scores take non-negative values, often in $[0, 1]$, e.g., using iForest [18], (ii) *non-decreasing* (see Lemma 4) and (iii) *submodular* (see Lemma 5).

Lemma 4 (Monotonicity). *f is non-decreasing, i.e., for any $\mathcal{S} \subseteq \mathcal{T}$, $f(\mathcal{S}) \leq f(\mathcal{T})$.*

Proof. $f(\mathcal{S}) = \sum_{a_i \in \mathcal{A}} \max_{p_j \in \mathcal{S}} s_{i,j} \leq \sum_{a_i \in \mathcal{A}} \max_{p_j \in \mathcal{T}} s_{i,j} = f(\mathcal{T})$ □

Lemma 5 (Submodularity). *f is submodular, i.e., for any two sets $\mathcal{S} \subseteq \mathcal{T}$ and a focus-plot $p_{j^*} \in \mathcal{P} \setminus \mathcal{T}$, $f(\mathcal{S} \cup \{p_{j^*}\}) - f(\mathcal{S}) \geq f(\mathcal{T} \cup \{p_{j^*}\}) - f(\mathcal{T})$.*

Proof

$$
\begin{aligned}
f(\mathcal{S} \cup \{p_{j^*}\}) - f(\mathcal{S}) &= \sum_{a_i \in \mathcal{A}} \left[\max_{p_j \in \mathcal{S} \cup \{p_{j^*}\}} s_{i,j} - \max_{p_j \in \mathcal{S}} s_{i,j} \right] \\
&= \sum_{a_i \in \mathcal{A}} \left(s_{i,j^*} - \max_{p_j \in \mathcal{S}} s_{i,j} \right) \cdot \mathbb{I}\left[s_{i,j^*} > \max_{p_j \in \mathcal{S}} s_{i,j} \right] \\
&\geq \sum_{a_i \in \mathcal{A}} \left(s_{i,j^*} - \max_{p_j \in \mathcal{T}} s_{i,j} \right) \cdot \mathbb{I}\left[s_{i,j^*} > \max_{p_j \in \mathcal{S}} s_{i,j} \right] \qquad (2) \\
&\geq \sum_{a_i \in \mathcal{A}} \left(s_{i,j^*} - \max_{p_j \in \mathcal{T}} s_{i,j} \right) \cdot \mathbb{I}\left[s_{i,j^*} > \max_{p_j \in \mathcal{T}} s_{i,j} \right] \qquad (3) \\
&= f(\mathcal{T} \cup \{p_{j^*}\}) - f(\mathcal{T})
\end{aligned}
$$

where $\mathbb{I}[\cdot]$ is the indicator function and Eqs. (2) and (3) follow from the fact that $\max_{p_j \in \mathcal{S}} s_{i,j} \leq \max_{p_j \in \mathcal{T}} s_{i,j}$ whenever $\mathcal{S} \subseteq \mathcal{T}$. □

Proposed LookOut Algorithm. Submodular functions which are non-negative and non-decreasing admit approximation guarantees under a greedy approach identified by Nemhauser et al. [21]. The greedy algorithm starts with the empty set $\mathcal{S}_0$. In iteration t, it adds the element (in our case, focus-plot) that maximizes the *marginal gain* Δ_f in function value, defined as

$$\Delta_f(p|\mathcal{S}_{t-1}) = f(\mathcal{S}_{t-1} \cup \{p\}) - f(\mathcal{S}_{t-1}) \tag{4}$$

That is,

$$\mathcal{S}_t := \mathcal{S}_{t-1} \cup \{ \operatorname*{arg\,max}_{p \in \mathcal{P} \setminus \mathcal{S}_{t-1}} \Delta_f(p|\mathcal{S}_{t-1}) \} .$$

This leads to LookOut explanation algorithm, given in Algorithm 1. Its approximation guarantee is given in Lemma 6.

Lemma 6 (63% approximation guarantee). *Given $\mathcal{A}, \mathcal{P}$ and budget b, let $\hat{\mathcal{S}}$ be the output of* LookOut *(Algorithm 1). Suppose $\mathcal{S}^* = \arg\max_{\mathcal{S} \subseteq \mathcal{P}, |\mathcal{S}|=b} f(\mathcal{S})$ is an optimal set of focus-plots. Then:*

$$f(\hat{\mathcal{S}}) \geq \left(1 - \frac{1}{e}\right) f(\mathcal{S}^*) \tag{5}$$

Proof. This follows from [21] since by design, our plot selection objective f is non-negative, non-decreasing and submodular. □

4.4 Computational Complexity Analysis

Lemma 7. LookOut *total time complexity is $O(l \log n'(k+n') + klb)$, for sample size $n' < n$, and is sub-linear in total number of input points n.*

Proof. We study complexity in two parts: (1) scoring the given outliers (Sect. 4.1) and (2) selecting focus-plots to present to the user (Sect. 4.2).

(1) For each focus-plot, we train an iForest model [18] in 2-d. Following their recommended setup, we sub-sample n' points and train t (100 in [18]) randomized isolation trees. The depth of each tree is $O(\log n')$, where each point is evaluated at each level for the threshold/split conditions. Therefore, training iForest with t trees takes $O(tn' \log n')$. Then, scoring $|\mathcal{A}| = k$ outliers takes $O(tk \log n')$. Total complexity of training and scoring on all plots is $O(lt \log n'(k + n'))$. Note that this can also be done per plot independently in parallel to reduce time.
(2) At each iteration of the greedy selection algorithm, we compute the marginal gain for each yet-unselected plot of being added to our select-set in $O(kl)$. Marginal gain per plot can also be computed independently in parallel. Among the remaining plots, we pick the one with the largest marginal gain. Finding the maximum among all gains takes $O(l)$ via a linear scan. We repeat this process b times until the budget is exhausted. Total selection complexity is thus $O(klb)$.

The overall complexity of both parts is effectively $O(l \log n'(k + n') + klb)$, since t is a constant. □

Notice that the total number of focus-plots, $l = d^2$, is quadratic in number of features. Typically, d is small (<100). In high dimensions, we could either use parallelism (multi-core machines are commodity), or drop features with low kurtosis as done earlier [18] or other feature selection criteria [13].

4.5 Discussion

Here we answer some questions that may be in the reader's mind.

1. *How do we define "outlier?"* We defer this question to the off-the-shelf outlier detection algorithm (iForest [18], LOF [5], etc.). Our focus here is to succinctly and interpretably *explain* what makes the selected items stand out from the rest.
2. *Why focus-plots?* Using focus-plots for justification is an essential, conscious choice we make for several reasons: (a) scatter plots are easy to look at and quickly interpret (b) they are universal and non-verbal, in that we need not use language to convey the outlierness of points – even people unfamiliar with the context of Enron will agree that the point "Jeff Skilling" in Fig. 1 is far away from the rest, and (c) they show where the outliers lie *relative to the normal points* – the contrastive visualization of points is more convincing than stand-alone rules.
3. *How do we choose the budget b?* We designed our objective function to be budget-conscious, and let the budget be specified by the analyst (user). If not specified, we use $b = 7$, since humans have a working memory of size "seven, plus or minus two" [20].
4. *Why not decision trees to separate outliers from the rest?* While arguably interpretable, decisions trees are not easy to visualize the points when higher than depth 3. Moreover, they try to find balanced splits which would try to cluster the outliers – which is unlikely for outliers. Also, decision trees are not budget-conscious, i.e. they would not necessarily produce the minimum description. Finally, they do not provide any quantifiable explanations, i.e. incrimination per outlier like our $s_{i,j}$ scores – the splits are binary.

5 Experiments

In this section, we empirically evaluate LookOut on three, diverse datasets. Our experiments were designed to answer the following questions:

[Q1] Quality of Explanation: How well can LookOut "explain" or "blame" the given outliers?

[Q2] Scalability: How does LookOut scale with the input graph size and the number of outliers?

[Q3] Discoveries: Does LookOut lead to interesting and intuitive explanations on real world data?

These are addressed in Sects. 5.3, 5.4 and 5.5 respectively. Before detailing our empirical findings, we describe the datasets used and our experimental setup.

5.1 Dataset Description

To illustrate the generality of our proposed domain-agnostic pictorial outlier explanations algorithm LookOut, we select our datasets from diverse domains: - e-mail communication (Enron), co-authorship (Dblp), pulsar identification (Htru), and glass composition (Glass). All datasets are publicly available and the first two are unipartite, directed and undirected resp., time-evolving graph datasets. The latter two are multi-feature datasets consisting of continuous values. A brief description is given below and a summary is provided in Table 2.

Table 2. Datasets with labeled outliers (that we explain) studied in this work

Dataset	Type	# points	# features	Description
Enron[a]	Graph	151	12	e-mail communications
Dblp[b]	Graph	1.3M	12	Co-authorship
Htru[c]	Feature	17.9K	8	Pulsar identification
Glass[d]	Feature	213	9	Glass composition

[a] http://networkdata.ics.uci.edu/netdata/html/EnronMailUSC1.html
[b] http://konect.uni-koblenz.de/networks/dblp_coauthor
[c] https://archive.ics.uci.edu/ml/datasets/Glass+Identification
[d] https://archive.ics.uci.edu/ml/datasets/HTRU2

Enron: This dataset consists of 19K emails exchanged between 151 Enron employees during the period surrounding the scandal[5] (May 1999–June 2002).
Dblp: This dataset contains the co-authorship network of 1.3M authors over 25 years from 1990 to 2014. The networks are collected at yearly granularity.
Htru: This dataset describes a sample of 17.9K pulsar (rapidly rotating neutron star) candidates collected during the High Time Resolution Universe Survey. Radio emissions have been binary classified as either background noise or as pulsar radiation. Features are extracted from the radio emission pattern curves.
Glass: This dataset consists of a multi-class classification of 213 glass samples with element-wise compositions of each sample as features. There are a total of seven classes which are clustered into two distinct types: (Classes 1–4) window glass and (Classes 5–7) non-window glass.

[5] https://en.wikipedia.org/wiki/Enron_scandal.

5.2 Experimental Setup

Graph Feature Extraction: We extract the following intuitive and easy-to-understand features from our graph datasets (Enron, Dblp) in order to generate pictorial explanations: (1) `indegree` and (2) `outdegree` for the number of unique in- and out- neighbors of every node, (3) `inweight-v` and (4) `outweight-v` for the total weight of in- and out- edges incident on each node, (5) `inweight-r` and (6) `outweight-r` for the count of in- and out- edges (including repetitions) incident on each node, (7) `average-IAT`, (8) `IAT-variance`, (9–11) `minimum-IAT`, `median-IAT`, and `maximum-IAT` to capture various statistics of inter-arrival time (IAT) between edges and finally, (12) `lifetime`- for the time gap between the first and the last edge [2,11,22].

Groundtruth: To obtain "ground-truth" outliers for LookOut input, we use the iForest [18] algorithm on given or extracted features. This yields a ranked list of points with scores in $[0, 1]$ (higher value suggests higher abnormality), from which we pick the desired top k. Analogously, we use iForest for computing the outlier score in each focus-plot. We note that the analyst is free to choose any outlier detector(s) for both/either of the above stages, making LookOut *detector-agnostic*. However, it is recommended that the same methods be used for both stages to ensure ranking similarities.

Evaluation Metric: We quantify the quality of explanation provided by a set of plots $\mathcal{S}$ using its *incrimination* score which is a normalized form of our objective:

$$\textit{incrimination}(\mathcal{S}) = \frac{1}{C} \cdot f(\mathcal{S}) \tag{6}$$

where C is the normalization constant equal to the maximum achievable objective (see Eq. (1)) when all plots are selected, i.e., $C = f(\mathcal{P})$.

Baselines: Due to the lack of comparable prior works, we use a naïve version of our approach, called LookOut-Naïve which ignores the submodularity of our objective. Instead, LookOut-Naïve assigns a score to each plot by summing up scores for all given outliers and chooses the top b plots for a given budget b. For the sake of comparison, we compare both LookOut and LookOut-Naïve with a Random baseline in which random b plots are chosen for the given budget b.

All experiments were performed on an OSX computer with 16 GB memory. Random baseline incriminations and runtimes were averaged over 10 trials.

5.3 Quality of Explanation

Figure 4 compares the *incrimination* scores of both LookOut-Naïve and LookOut on the Enron, Htru and Glass datasets for several choices of k and b. The red dotted line indicates the ideal value, $\textit{incrimination}(\mathcal{P}) = 1$, i.e., the highest achievable incrimination (by selecting all plots). Figure 4 shows that LookOut consistently outperforms the baselines and rapidly converges to the ideal *incrimination* with increasing budget.

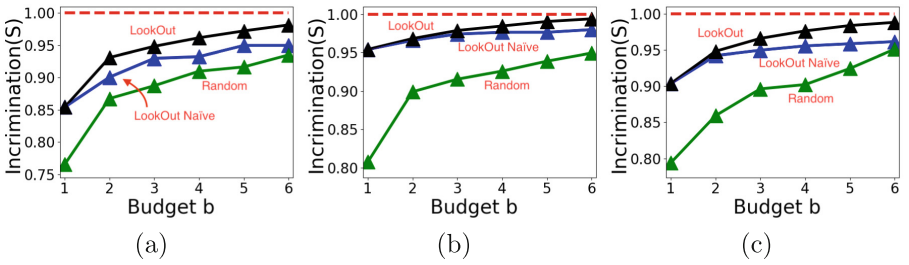


Fig. 4. LookOut vs. baselines: (a) Enron with $k = 10$ outliers, (b) Htru with $k = 32$ outliers and (c) Glass with $k = 28$ outliers (Color figure online)

5.4 Scalability

We empirically studied how LookOut runtime varies with (i) number of focus-plots l and (ii) the number of outliers k.

To study the variation of runtime with the number of focus-plots, we vary the number of features which are taken into consideration. Figure 5 (left) illustrates linear scaling with respect to number of focus-plots for the Glass dataset.

We also study the variation of runtime with the number of outliers, as feature extraction incurs a constant overhead on each dataset. Figure 5 (right) shows linear scaling with the number of outliers for a Dblp subgraph with 10K edges.

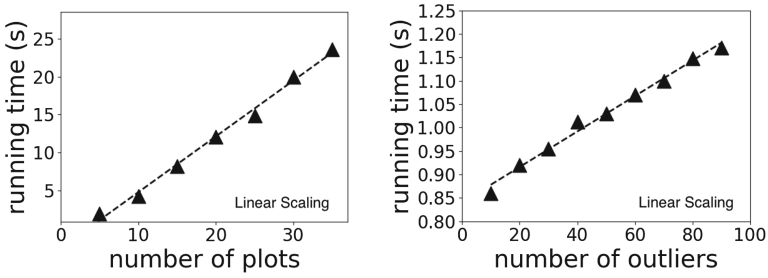


Fig. 5. LookOut **scales linearly** with (left) number of focus-plots to consider and right) number of outliers

5.5 Discoveries

In this section, we present our discoveries using LookOut on all four real world datasets. Scoring in 2-d was performed using iForest with $t = 100$ trees and sample size $\psi = 64$ (Enron, Htru, Glass) and $\psi = 256$ (Dblp). We use dictated outliers for Enron, Htru and Glass, and detected outliers for Dblp dataset to demonstrate performance in both settings.

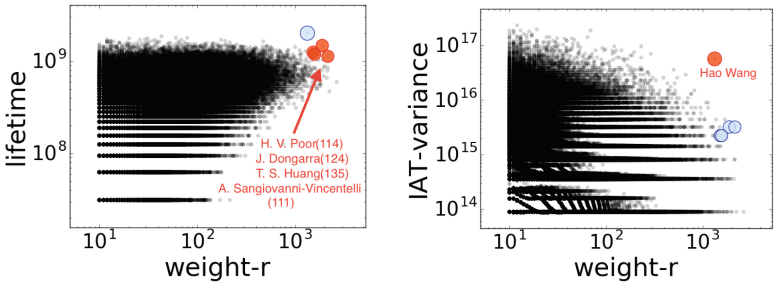


Fig. 6. Discoveries using LookOut on detected outliers: LookOut partitions and explains outlier detection results from iForest on Dblp (a–b)

Enron *(CEO & CFO explained by large out-degree)*
We used two top actors in the Enron scandal, *Kenneth Lay* (CEO) and *Jeff Skilling* (CFO) as dictated outliers for LookOut and sought explanations for their abnormality based on internal e-mail communications. With $b = 2$, LookOut produced the focus-plots shown in Fig. 1 (right). Explanations indicate that *Jeff Skilling* had an unusually large `IAT-max` for the number of employees he communicated with (`outdegree`). On the other hand, *Kenneth Lay* sent emails to an abnormally large number of employees (`outdegree`) given the time range during which he emailed anyone (`lifetime`).

Dblp *(high h-index authors explained by large lifetime and high co-authorships)*
We obtained ground truth outliers by running iForest on the high-dimensional space spanned by the extracted graph features. With $k = 5$, the detected outlying authors were *Jack Dongarra*, *Thomas S. Huang*, *Alberto L. Sangiovanni-Vincentelli*, *H. Vincent Poor*, and *Hao Wang*. The explanations provided by LookOut with $b = 2$ are shown in Fig. 6a–b. Thus, the outlying authors users are partitioned into two groups. The members of the first group, *Jack Dongarra*, *Thomas S. Huang*, *Alberto L. Sangiovanni-Vincentelli*, and *H. Vincent Poor* are outlying because they had unusually high duration during which they published papers (`lifetime`) and total number of co-authorships (`inweight-r`). This is consistent with their high h-indices obtained from their respective Google scholar pages (see brackets in Fig. 6a). The second group consists of only *Hao Wang*, who was also outlying in the first focus-plot, but is best explained by very high `IAT-variance` for his `inweight-r` value, shown in Fig. 6b.

Htru *(pulsars correlated with skewness and extra kurtosis of integrated profile)*
The given radio emission samples were classified as either random noise or pulsar generated. We subsampled datapoints from the pulsar class and considered them as our set of dictated outliers with $k = 32$. We ran LookOut on this subsampled dataset with $b = 3$ and obtained the focus-plots shown in Fig. 7a–c. The explanations infer high values of `skewness` and `excess kurtosis` strongly

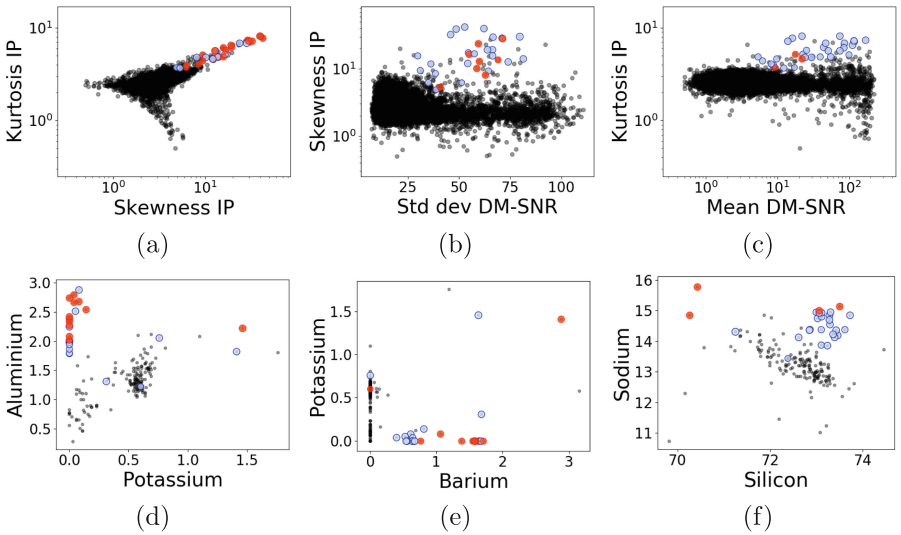


Fig. 7. Discoveries using LookOut on dictated outliers: LookOut explains outlier characteristics on Htru (a–c) and Glass (d–f) for budget $b = 3$.

indicate a pulsar emission. We observe that only the first focus-plot, Fig. 7a, succeeds to provide a suitable explanation for our set of detected anomalies. This is quantitatively explained in Fig. 4c where budget $b = 1$ has high incrimination and on further increasing the budget only a small marginal gain is observed.

Glass *(headlamp glass explained by high Aluminium and Barium content)* The dataset contains seven classes pertaining to different types of glass. Broadly these seven classes are split into two categories: *window based glass* (class 1–4) and *non-window based glass* (class 5–7). To compare glass composition between these two categories we stitched together a subset of the original dataset by including only classes 1, 2, 3, 4 & 7. Here class 7 (*headlamps*) is considered the set of dictated outliers with $k = 28$. The explanations provided by LookOut, on the newly constructed dataset, with $b = 3$ are shown in Fig. 7d–f. The first two focus-plots Fig. 7d–e reflect higher `aluminium` and `barium` concentrations in *headlamps* as compared to *window glass*. `Aluminium` is used as a reflective coating and the presence of `barium`, in the form of oxides (borosilicate glass), helps induce heat resistant properties – both properties we expect to find in headlamps. Concurrently, we observe a very low or nearly zero concentration of `potassium` in headlamp glass. `Potassium` is to used to toughen glass and is found in windows which need to be resistant to adverse weather conditions.

Note that on all datasets, outlying points are visually distinguishable, and often complementary between focus-plots. This is in line with our desired explanation task, and achieved as a result of our LookOut subset selection objective.

6 Conclusions

In this work, we formulated and tackled the problem of succinctly and interpretably explaining outliers to human analysts. We made the following contributions: (a) **problem formulation:** we formulate our goal for explaining outliers using a budget of visually interpretable focus-plots, (b) **explanation algorithm:** we propose a submodular objective to quantify explanation quality and propose the LookOut method for solving it approximately with guarantees, (c) **generality:** we show that LookOut can work with diverse domains and any detection algorithm, and (d) **scalability:** we show theoretically and empirically that LookOut scales linearly in the number of input outliers as well as the total number of focus-plots to chose from. We conduct experiments on real-world datasets: e-mail communication, co-authorship, pulsar identification, and glass composition and demonstrate that LookOut produces qualitatively interpretable explanations for "ground-truth" outliers and achieves strong quantitative performance in maximizing our proposed objective.

Acknowledgments. This material is based upon work supported by the National Science Foundation (NSF) under Grants No. CNS-1314632, IIS-1408924 and IIS 1408287, by NSF CAREER 1452425 and the PwC Risk and Regulatory Services Innovation Center at Carnegie Mellon University. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation, or other funding parties. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

1. Aggarwal, C.C.: Outlier Analysis. Springer, New York (2013). https://doi.org/10.1007/978-1-4614-6396-2
2. Akoglu, L., McGlohon, M., Faloutsos, C.: oddball: spotting anomalies in weighted graphs. In: Zaki, M.J., Yu, J.X., Ravindran, B., Pudi, V. (eds.) PAKDD 2010. LNCS, vol. 6119, pp. 410–421. Springer, Heidelberg (2010). https://doi.org/10.1007/978-3-642-13672-6_40
3. Akoglu, L., Tong, H., Koutra, D.: Graph based anomaly detection and description: a survey. Data Min. Knowl. Discov. **29**(3), 626–688 (2015)
4. Angiulli, F., Fassetti, F., Palopoli, L.: Discovering characterizations of the behavior of anomalous subpopulations. IEEE TKDE **25**(6), 1280–1292 (2013)
5. Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: LOF: identifying density-based local outliers. SIGMOD Rec. **29**(2), 93–104 (2000)
6. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection: a survey. ACM Comput. Surv. (CSUR) **41**(3), 15 (2009)
7. Dang, X.H., Assent, I., Ng, R.T., Zimek, A., Schubert, E.: Discriminative features for identifying and interpreting outliers. In: ICDE, pp. 88–99 (2014)
8. Dang, X.H., Micenková, B., Assent, I., Ng, R.T.: Local outlier detection with interpretation. In: Blockeel, H., Kersting, K., Nijssen, S., Železný, F. (eds.) ECML PKDD 2013. LNCS, vol. 8190, pp. 304–320. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-40994-3_20

9. Emmott, A.F., Das, S., Dietterich, T., Fern, A., Wong, W.K.: Systematic construction of anomaly detection benchmarks from real data. In: KDD Workshop on Outlier Detection and Description (2013)
10. Garey, M.R., Johnson, D.S.: Computers and Intractability: A Guide to the Theory of NP-Completeness. Freeman, New York (1979)
11. Giatsoglou, M., Chatzakou, D., Shah, N., Beutel, A., Faloutsos, C., Vakali, A.: ND-Sync: detecting synchronized fraud activities. In: Cao, T., Lim, E.-P., Zhou, Z.-H., Ho, T.-B., Cheung, D., Motoda, H. (eds.) PAKDD 2015. LNCS, vol. 9078, pp. 201–214. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-18032-8_16
12. Gupta, M., Gao, J., Aggarwal, C.C., Han, J.: Outlier detection for temporal data: a survey. IEEE Trans. Knowl. Data Eng. **26**(9), 2250–2267 (2014)
13. Traina Jr., C., Traina, A.J.M., Wu, L., Faloutsos, C.: Fast feature selection using fractal dimension. JIDM **1**(1), 3–16 (2010)
14. Keller, F., Müller, E., Wixler, A., Böhm, K.: Flexible and adaptive subspace search for outlier analysis. In: CIKM, pp. 1381–1390. ACM (2013)
15. Knorr, E.M., Ng, R.T.: Finding intensional knowledge of distance-based outliers. In: VLDB, pp. 211–222 (1999)
16. Kopp, M., Pevný, T., Holena, M.: Interpreting and clustering outliers with sapling random forests. In: ITAT (2014)
17. Kuo, C.T., Davidson, I.: A framework for outlier description using constraint programming. In: AAAI, pp. 1237–1243 (2016)
18. Liu, F.T., Ting, K.M., Zhou, Z.H.: Isolation forest. In: ICDM, pp. 413–422 (2008)
19. Micenková, B., Ng, R.T., Dang, X.H., Assent, I.: Explaining outliers by subspace separability. In: ICDM, pp. 518–527 (2013)
20. Miller, G.: The magic number seven plus or minus two: some limits on our automatization of cognitive skills. Psychol. Rev. **63**, 81–97 (1956)
21. Nemhauser, G.L., Wolsey, L.A.: Best algorithms for approximating the maximum of a submodular set function. Math. Oper. Res. **3**(3), 177–188 (1978)
22. Shah, N., et al.: EdgeCentric: anomaly detection in edge-attributed networks. In: ICDM Workshops, pp. 327–334. IEEE (2016)