Graph-Based Anomaly Detection: Problems, Algorithms and Applications

Leman Akoglu
Anomaly: That stands out

https://en.wikipedia.org/wiki/August_Landmesser
Anomaly Detection: Many Use-cases
Formalizing Anomaly Detection

• Concrete problem settings exist. e.g./esp. for point-cloud data

• Real-world… A bit more complex.
Formalizing Anomaly Detection

Given <DATA>, Find <ANOMALIES>

e.g. (accounting) Given millions of transactions, find abnormalities

We heard you work on anomaly detection.

Yes, I am very excited. Tell me more.

We have lots of data, and want to find anomalies.

OK, wait, tell me what your REAL PROBLEMS are.

Why do you want to detect anomalies?

What do you consider to be an anomaly?
Tell me what your REAL PROBLEMS are.

We want to find errors, inefficiencies, malfeasance… We want to save $$$$. We also want YOU to find all unknown anomalies.

Hmm… OK…?
Graph-based Anomaly Detection

- Often, underlying data is unmistakably **relational**:

  - user-business reviews
  - employer-employee
  - account transactions
  - physician-patient-provider
Graph-based Anomaly Detection

Several surveys and tutorials:

[Survey] Graph-based Anomaly Detection and Description: A Survey. [Akoglu+]
Data Mining and Knowledge Discovery (DAMI), May 2015.

[Tutorial] Fraud Detection through Graph-Based User Behavior Modeling. [Beutel+]
ACM CCS 2015.

[Tutorial] Social Media Anomaly Detection: Challenges and Solutions. [Liu & Chawla]
ACM SIGKDD 2015.

[Survey] False Information on Web and Social Media: A Survey. [Kumar & Shah]
arXiv:1804.08559
Challenges

**Problem:** Given \(<\text{Data}>\), Find \(<\text{Anomalies}>\) s.t. \(<\text{Constraints}>\)

1. \(<\text{Data}>\) : Graph **heterogeneity** (node/edge labels, attributes, multi-edges, edge weights, edge timestamps, etc.)
   How or whether to “fold” **meta-data** into a graph

2. \(<\text{Anomalies}>\) : Definition/Formalization of anomalies (e.g., group anomalies vs. anomalous groups)
   Heterogeneity exacerbates the issue

3. \(<\text{Constraints}>\) : System/Application **requirements** e.g., distributed/streaming/massive data, attribution (who), explainability (why)
Outline

• Anomaly Detection: Motivation, Formalism, Challenges

• Graph-based Anomaly Detection
  • General-purpose (single graph)
    Global – anomalous nodes
    Local – group anomalies
    Collective – anomalous groups
  • Specialized (graph database)

• Recent Trend: Deep Anomaly Detection
Anomalous nodes (global)

Problem Sketch:

plain, weighted, directed
Anomalous nodes (global): OddBall

Problem Setting:

- For each node
  - Extract ego-net (1-hop neighborhood)
  - Extract ego-net features
- Find patterns ("laws")
- Detect outliers (distance to patterns)

Anomalous nodes (global): OddBall

discussion group, “rank boosting”, etc.

telemarketer, spammer, port scanner, “popularity contests”, etc.
Anomalous nodes (global): OddBall

FORUM POSTS

http://www.sizemore.co.uk/2005/08/i-feel-some-movies-coming-on.html


#cross-citations vs. #citations
Anomalous nodes (global): OddBall

CAMPAIGN-DONATIONS

Kerry, John F.

Snyder, James E. Jr

Russo, Aaron

L. Akoglu
Outline

• Anomaly Detection: Motivation, Formalism, Challenges

• Graph-based Anomaly Detection
  • General-purpose (single graph)
    Global – anomalous nodes
    Local – group anomalies
    Collective – anomalous groups
  • Specialized (graph database)

• Recent Trend: Deep Anomaly Detection
Group anomalies (local)

Problem Sketch:

(left) community “focuses” on \{\text{degree}, \text{location}\} (right) “focuses” on work.
Group anomalies (local): FocusCO

cluster extraction & (cluster) outlier detection

1. similar pairs
2. dissimilar pairs
3. "focus" estimation

age
gender
Group anomalies (local): FocusCO

- **Focus** estimation

- Clusters & **Local** outliers

S and D (intermixed)
Group anomalies (local): FocusCO

- **Focus** estimation

  - **Clusters & Local outliers**

  (Local) clustering obj.: conductance $\phi^{(w)}$
  
  weighted by **focus**

  $\phi^{(w)}(C, G) = \frac{W_{cut}(C)}{W_{Vol}(C)}$

  Cluster Member

  Focused Outlier
  
  node with many (but) weak ties

  Feature Matrix

  S and D (intermixed)
Group anomalies (local): FocusCO

Cluster Outlier did not mention Waas.

Liberal Cluster in Political Blogs Graph
Outline

• Anomaly Detection: Motivation, Formalism, Challenges

• Graph-based Anomaly Detection
  • General-purpose (single graph)
    Global – anomalous nodes
    Local – group anomalies
    Collective – anomalous groups
  • Specialized (graph database)

• Recent Trend: Deep Anomaly Detection
Anomalous groups (collective)

• Problem Sketch:
Given a node-attributed subgraph, how to define “normality”?
Anomalous groups (collective): AMEN

Problem Setting:

- Given a subgraph in a node-attributed graph
- Identify (subgraph) “focus” such that
  - Internal nodes are structurally dense & coherent in focus
  - External nodes are structurally sparse or not-surprising, or different in focus

Anomalous groups (collective): AMEN

- Internal nodes are structurally dense & coherent in focus
- External nodes are structurally sparse or not-surprising, or different in focus
Anomalous groups (collective): AMEN

- Measure of Normality:

\[ N = I + E = \sum_{i \in C, j \in C} (A_{ij} - \frac{k_i k_j}{2m}) s(x_i, x_j | w) \]

\[ - \sum_{i \in C, b \in B \atop (i, b) \in E} (1 - \min(1, \frac{k_i k_b}{2m})) s(x_i, x_b | w) \]
Anomalous groups (collective): AMEN

- Estimating Normality:

\[
N = I + E = \sum_{i \in C, j \in C} (A_{ij} - \frac{k_i k_j}{2m}) s(x_i, x_j | w)
- \sum_{i \in C, b \in B, (i, b) \in E} (1 - \min(1, \frac{k_i k_b}{2m})) s(x_i, x_b | w)
\]

\[
\max_{w_C} w_C^T \cdot \left[ \sum_{i \in C, j \in C} (A_{ij} - \frac{k_i k_j}{2m}) s(x_i, x_j) 
- \sum_{i \in C, b \in B, (i, b) \in E} (1 - \min(1, \frac{k_i k_b}{2m})) s(x_i, x_b) \right]
\]

\[
\max_{w_C} w_C^T \cdot (\hat{x}_I + \hat{x}_E)
\]

s.t. \[\|w_C\|_p = 1, \ w_C(f) \geq 0, \forall f = 1 \ldots d\]
Anomalous groups (collective): AMEN

- Telescopic op-amps
- Split-radix FFT

Telescopic cascode multiphase

Reciprocal reserve split radix
Anomalous groups (collective): AMEN

- Telescopic op-amps
- Split-radix FFT

DBLP
$L_1 = 0.979, L_2 = 2.17$

Twitter
$L_1 = 0.724, L_2 = 1.10$

Citeeseer
$L_1 = L_2 = -0.956$

Google+
$L_1 = L_2 = -0.873$
Anomalous groups in malice detection

- In other contexts, “too dense”ly connected groups may be indicative of malice/fraud

- 9/11 hijackers were densely linked via
  - kinship
  - school/training
  - travel/financial records
  - meetings
  - …
Anomalous groups in malice detection

- **Opinion fraud**: Groups of users promoting/demoting businesses

<table>
<thead>
<tr>
<th>Prior</th>
<th>Review Ranking</th>
<th>AP</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Y’Chi</td>
<td>Y’NYC</td>
</tr>
<tr>
<td>Random</td>
<td>0.1327</td>
<td>0.1028</td>
<td>0.1321</td>
</tr>
<tr>
<td>FraudEagle</td>
<td>0.1067</td>
<td>0.1122</td>
<td>0.1524</td>
</tr>
<tr>
<td>Wang et al.</td>
<td>0.1518</td>
<td>0.1255</td>
<td>0.1803</td>
</tr>
<tr>
<td>Prior</td>
<td>0.2241</td>
<td>0.1789</td>
<td>0.2352</td>
</tr>
<tr>
<td>Speagle</td>
<td><strong>0.3236</strong></td>
<td><strong>0.2460</strong></td>
<td><strong>0.3319</strong></td>
</tr>
</tbody>
</table>

BIRDNEST: Bayesian Inference for Ratings-Fraud Detection. [Hooi+] SIAM SDM 2016
Anomalous groups in malice detection

• **Social securities tax fraud**: Groups of resources transferred between “shadow” companies

Outline

• Anomaly Detection: Motivation, Formalism, Challenges

• Graph-based Anomaly Detection
  • General-purpose (single graph)
    Global – anomalous nodes
    Local – group anomalies
    Collective – anomalous groups
  • Specialized (graph database)

• Recent Trend: Deep Anomaly Detection
Anomalous graphs (System security)

- Advanced Persistent Threat

Problem Setting:

- **Given** a stream of event logs
- **Find** anomalous system events

<table>
<thead>
<tr>
<th>time</th>
<th>pid</th>
<th>event</th>
<th>arg/data</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>10639</td>
<td>fork</td>
<td>NULL</td>
</tr>
<tr>
<td>200</td>
<td>10640</td>
<td>execve</td>
<td>/bin/sh</td>
</tr>
<tr>
<td>300</td>
<td>10650</td>
<td>read</td>
<td>STDIN</td>
</tr>
<tr>
<td>400</td>
<td>10640</td>
<td>fstat</td>
<td>0xbfc5598</td>
</tr>
<tr>
<td>500</td>
<td>10660</td>
<td>sock_wr</td>
<td>0.0.0.0</td>
</tr>
</tbody>
</table>

---

Host-level detection
Anomalous graphs (System security)

- Advanced Persistent Threat

Problem Setting:

- **Given** a stream of event logs
- **Find** anomalous system events

Given `<DATA>`, Find `<ANOMALIES>` s.t. `<CONSTRAINTS>`

Requirements:

- Real-time detection
- Low-latency
- Low computational overhead
- Low memory usage
Anomalous graphs (System security)

- Each event associated with a logical flow (tag)

<table>
<thead>
<tr>
<th>time</th>
<th>pid</th>
<th>event</th>
<th>arg/data</th>
<th>tag</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>10639</td>
<td>fork</td>
<td>NULL</td>
<td>1</td>
</tr>
<tr>
<td>200</td>
<td>10640</td>
<td>execve</td>
<td>/bin/sh</td>
<td>1</td>
</tr>
<tr>
<td>300</td>
<td>10650</td>
<td>read</td>
<td>STDIN</td>
<td>2</td>
</tr>
<tr>
<td>400</td>
<td>10640</td>
<td>fstat</td>
<td>0xbfc5598</td>
<td>1</td>
</tr>
<tr>
<td>500</td>
<td>10660</td>
<td>sock_wr</td>
<td>0.0.0.0</td>
<td>2</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

- Events from different flows may interleave

Each event as a directed edge:

- 10639 → <100, fork>
- 10650 → <500, sock_wr>
- 10640
- 10660
- STDIN

! Many, simultaneously-growing node&edge-labeled graphs
! Universe of labels unknown

Anomalous graphs (Accounting)

- **Double-entry Bookkeeping example journal entry:**

<table>
<thead>
<tr>
<th>GL_Account_Number</th>
<th>CA_FS_Caption</th>
<th>Cr/Db</th>
<th>GL_Reporting_Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>40020000 (Revenue)</td>
<td>Gross Sales (GSL)</td>
<td>C</td>
<td>-7250</td>
</tr>
<tr>
<td>40020001 (Revenue)</td>
<td>Gross Sales (GSL)</td>
<td>C</td>
<td>-2500</td>
</tr>
<tr>
<td>20830000 (Liabilities)</td>
<td>Sales Tax Payables (STP)</td>
<td>C</td>
<td>-794.63</td>
</tr>
<tr>
<td>10390000 (Assets)</td>
<td>Accounts Receivable (ARV)</td>
<td>D</td>
<td>10544.63</td>
</tr>
</tbody>
</table>

**Problem Setting:**

- **Given** millions of journal entries
- **Find** anomalies

(entry errors, misconduct, etc.)

Given <DATA>, Find <ANOMALIES> s.t. <CONSTRAINTS>
Anomalous graphs (Accounting)

- Double-entry Bookkeeping example journal entry:

<table>
<thead>
<tr>
<th>GL_Account_Number</th>
<th>CA_FS_Caption</th>
<th>Cr/Db</th>
<th>GL_Reporting_Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>40020000 (Revenue)</td>
<td>Gross Sales (GSL)</td>
<td>C</td>
<td>-7250</td>
</tr>
<tr>
<td>40020001 (Revenue)</td>
<td>Gross Sales (GSL)</td>
<td>C</td>
<td>-2500</td>
</tr>
<tr>
<td>20830000 (Liabilities)</td>
<td>Sales Tax Payables (STP)</td>
<td>C</td>
<td>-794.63</td>
</tr>
<tr>
<td>10390000 (Assets)</td>
<td>Accounts Receivable (ARV)</td>
<td>D</td>
<td>10544.63</td>
</tr>
</tbody>
</table>

Problem Setting:
- **Given** millions of journal entries
- **Find** anomalies (entry errors, misconduct, etc.)

Given <DATA>, Find <ANOMALIES> s.t. <CONSTRAINTS>

Requirements:
- Explainability (audit)

Anomalous graphs (Accounting)

- **Transaction graphs:**

<table>
<thead>
<tr>
<th>GL_Account_Number</th>
<th>CA_FS_Caption</th>
<th>Cr/Db</th>
<th>GL_Reporting_Amount</th>
</tr>
</thead>
<tbody>
<tr>
<td>40060000</td>
<td>Gross Sales (GSL)</td>
<td>C</td>
<td>-1575.00</td>
</tr>
<tr>
<td>10415000</td>
<td>Accounts Receivable (ARV)</td>
<td>D</td>
<td>1575.00</td>
</tr>
<tr>
<td>40020000</td>
<td>Gross Sales (GSL)</td>
<td>C</td>
<td>-7250</td>
</tr>
<tr>
<td>40020001</td>
<td>Gross Sales (GSL)</td>
<td>C</td>
<td>-2500</td>
</tr>
<tr>
<td>20830000</td>
<td>Sales Tax Payables (STP)</td>
<td>C</td>
<td>-794.63</td>
</tr>
<tr>
<td>10390000</td>
<td>Accounts Receivable (ARV)</td>
<td>D</td>
<td>10544.63</td>
</tr>
</tbody>
</table>
Anomalous graphs (Accounting)

- Transaction graph of journals over 10-day window:
Anomalous graphs (Accounting)

- Anomaly detection via data description/encoding

\[
\min_{MT \subseteq M} L(MT, G) = \underbrace{L(MT)}_{\text{model code length}} + \underbrace{L(G|MT)}_{\text{data code length}}, \quad (1)
\]

Graph cover \( G_j \)

\[ MT \]

\[
g_1 \quad \begin{array}{c}
A \\
B \quad C
\end{array}
\quad c_1
\]

\[
g_2 \quad \begin{array}{c}
B \\
D \quad A
\end{array}
\quad c_2
\]

\[
g_3 \quad \begin{array}{c}
C \\
B
\end{array}
\quad c_3
\]

\cdots \quad \cdots
Anomalous graphs (Accounting)

- Anomaly detection via data description/encoding

\[
\text{minimize}_{MT \subseteq \mathcal{M}} \quad L(MT, G) = \underbrace{L(MT)}_{\text{model code length}} + \underbrace{L(G|MT)}_{\text{data code length}}, \quad (1)
\]
Anomalous graphs (Accounting)

- Anomaly detection via data description/encoding

<table>
<thead>
<tr>
<th>Method</th>
<th>Prec@10</th>
<th>Prec@100</th>
<th>Prec@1000</th>
<th>AUC</th>
<th>AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>CODETECT</td>
<td>0.900</td>
<td>0.990</td>
<td>0.999</td>
<td>0.995</td>
<td>0.772</td>
</tr>
<tr>
<td>SMT</td>
<td>0.600</td>
<td>0.440</td>
<td>0.784</td>
<td>0.906</td>
<td>0.733</td>
</tr>
<tr>
<td>SUBDUE</td>
<td>0.800</td>
<td>0.710</td>
<td>0.685</td>
<td>0.930</td>
<td>0.555</td>
</tr>
<tr>
<td>GF+iFOREST</td>
<td>0.400</td>
<td>0.230</td>
<td>0.497</td>
<td>0.959</td>
<td>0.429</td>
</tr>
<tr>
<td>G2V+iFOREST</td>
<td>0.000</td>
<td>0.100</td>
<td>0.819</td>
<td>0.824</td>
<td>0.380</td>
</tr>
<tr>
<td>DGK+iFOREST</td>
<td>0.300</td>
<td>0.140</td>
<td>0.023</td>
<td>0.858</td>
<td>0.097</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Prec@10</th>
<th>Prec@100</th>
<th>Prec@1000</th>
<th>AUC</th>
<th>AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>CODETECT</td>
<td>0.800</td>
<td>0.720</td>
<td>0.709</td>
<td>0.918</td>
<td>0.359</td>
</tr>
<tr>
<td>SMT</td>
<td>0.000</td>
<td>0.100</td>
<td>0.174</td>
<td>0.883</td>
<td>0.192</td>
</tr>
<tr>
<td>SUBDUE</td>
<td>0.800</td>
<td>0.710</td>
<td>0.685</td>
<td>0.920</td>
<td>0.555</td>
</tr>
<tr>
<td>GF+iFOREST</td>
<td>0.200</td>
<td>0.080</td>
<td>0.027</td>
<td>0.832</td>
<td>0.092</td>
</tr>
<tr>
<td>G2V+iFOREST</td>
<td>0.000</td>
<td>0.030</td>
<td>0.030</td>
<td>0.499</td>
<td>0.030</td>
</tr>
<tr>
<td>DGK+iFOREST</td>
<td>0.100</td>
<td>0.030</td>
<td>0.038</td>
<td>0.801</td>
<td>0.074</td>
</tr>
</tbody>
</table>
Outline

• Anomaly Detection: Motivation, Formalism, Challenges

• **Graph-based Anomaly Detection**
  - General-purpose (single graph)
    Global – anomalous nodes
    Local – group anomaly
    Collective – anomalous groups
  - Graph-level anomalies

Recent Trend: **Deep Anomaly Detection**
Deep Anomaly Detection

• Representation learning: transformative for applications in NLP/translation, recommender systems, etc.

• Why not automatically learn data representations for anomaly detection?


• Ideas easily transfer to graph data
Deep Anomaly Detection

Graph Embedding

\[ f(\text{[Karate Club Network]}) = \text{Outlier Detector} \]
Deep Anomaly Detection

Graph Embedding

- Can seamlessly handle various types of graphs: labeled, attributed, multi-edges, weights
- Can do end-to-end learning (one-class, reconstruction)
- Embeddings capture general prevalent patterns, may not be suitable for anomaly detection
- Hyper-parameter tuning becomes key for success!

Unsupervised model selection – likely a critical future direction
Graph-based Anomaly Detection

Code, Data, Papers, Slides

www.cs.cmu.edu/~lakoglu/

http://www.andrew.cmu.edu/user/lakoglu/pubs.html#code

Thanks!