An Embedding Approach to Anomaly Detection

Renjun Hu\textsuperscript{1}, Charu Aggarwal\textsuperscript{2}, Shuai Ma\textsuperscript{1}, and Jinpeng Huai\textsuperscript{1}

\textsuperscript{1}SKLSDE Lab, Beihang University, China
\textsuperscript{2}IBM T. J. Watson Research Center, USA
Motivation

- Anomaly detection
  - Identification of patterns in data that do not conform to expected behaviors [Chandola et al. 2009]
  - Useful in a wide variety of applications

- In networks, anomaly detection has broader meanings
  - Application-specific significance
  - Possibility to improve the performance of network-centric mining tasks such as community detection and classification

Motivation

- Structural hole theory [Burt 1992, 2004]
  - Theory of social capital
  - A structural hole is a gap between two nodes who have complementary sources to information

How to detect social brokers?
A formal quantitative definition is needed in the first place!

- Node A (social broker) is more likely to get novel information than B, even though they have the same number of links.

Motivation

- **Structural inconsistencies**
  - Nodes that connect to a number of diverse influential communities
  - Detect social brokers quantitatively

- **Anomalousness from homophily [McPherson et al. 2001]**
  - Linked nodes have similar properties
  - Fundamental to a wide variety of algorithms in network science
    - E.g., community detection, collective classification, link prediction, influence analysis
  - Violated by structural inconsistencies

---

Motivation

- **Structural inconsistencies**
  - Nodes that connect to a number of diverse influential communities
  - Detect social brokers quantitatively

- The presence of structural inconsistencies may:
  - have a substantial impact on network structure
    - *E.g.*, all nodes tend to form one large cluster
  - prevent effective applications of network mining algorithms
    - *E.g.*, hard for community detection algorithms to achieve meaningful clusters
Outline

- Anomaly detection model
  - Graph embedding
  - A quantitative measure of anomaly

- Algorithm optimization techniques

- Evaluation
Why graph embedding?

- Structural inconsistencies
  - connect to a number of diverse influential communities

- Evaluate the diversity or similarity of nodes. How?
  - To node B, node A is more similar than C, even though they have the same (global) distance from B.

- Graph embedding
  - Associate each node with a multidimensional vector
  - Preserve local linkage structure (instead of global structure)
  - Each dimension corresponds to a community in the network
Why graph embedding?

- Structural inconsistencies
  - connect to a number of diverse influential communities

- An alternative option: doing community detection followed by anomaly detection
  - Do not distinguish anomalies from normal nodes
  - The presence of anomalies has certain impacts on the results of community detection
  - Community detection is a heavy task.
  - Fail to detect structural inconsistencies!
Graph embedding

- Given an undirected graph $G=(V, E)$, associate each node $i$ with a $d$-dimensional vector $X_i$

- $V = \{1, 2, ..., n\}$
- $d$: number of communities
- $X_i$: correlation between node $i$ and the $d$ communities

A reasonable selection of $d$ suffices for anomaly detection. Not necessary to use the number of real-life communities.
Graph embedding

Given an undirected graph $G=(V, E)$, associate each node $i$ with a $d$-dimensional vector $X_i$

Goal: preserve local linkage structure

• Connected nodes should have similar values of $X_i$
• Disconnected nodes should have diverse values of $X_i$

Computation: minimizing objective function $O$

$$O = \sum_{(i,j) \in E} \|X_i - X_j\|^2 + \alpha \cdot \sum_{(i,j) \notin E} \left(1 - \|X_i - X_j\|\right)^2, \alpha = \frac{m}{\binom{n}{2} - m}$$

• $n$: number of nodes in $G$, $m$: number of edges in $G$
• $\alpha$: balancing factor that regulates the importance of the two components in $O$
• The embedding ensures that $0 \leq \|X_i - X_j\|^2 \leq 1$
A quantitative measure

- Inspired by structural inconsistencies and structural holes (social brokers)
  - Connect to a number of diverse influential communities
  - Bridge across complementary sources

- **NB(i):** how node $i$ connects to communities
  
  $$NB(i) = \left(y_i^1, ..., y_i^d\right) = \sum_{(i,j) \in E} \left(1 - \|X_i - X_j\|\right) \cdot X_j$$

- **AScore(i):** the anomalousness of node $i$

  $$AScore(i) = \sum_{k=1}^{d} \frac{y_i^k}{y_i^*}, y_i^* = \max\left\{y_i^1, ..., y_i^d\right\}$$

  - Detect anomalies by $AScore(i) > \text{thre}$
Example

- Optimality of embedding, *i.e.*, minimum value of $O$
  - Small values within groups because of missing edges
  - No values across groups
  - Certain values for the red node (no better embedding)

- Anomalousness of nodes
  - $\text{AScore(red)} = 4$ (equal values in dimensions of $NB(red)$)
  - $\text{AScore}(i) \approx 1$ for others ($NB(i)$ only has a dominating dimension)

\[
O = \sum_{(i,j)\in E} \|X_i - X_j\|^2 + \alpha \cdot \sum_{(i,j)\notin E} \left(1 - \|X_i - X_j\|^2 \right)^2
\]

\[
\text{AScore}(i) = \sum_{k=1}^{d} \frac{y_i^k}{y_i^*}, \quad y_i^* = \max \{ y_i^1, ..., y_i^d \}
\]

The red node is detected as an anomaly!
Outline

- Anomaly detection model

- Algorithm optimization techniques
  - Sampling
  - Graph partitioning based initialization
  - Dimension reduction

- Evaluation
Issues in the model

- Objective function $O$ is a sum over $O(n^2)$ terms
  - Forbidden in large social networks

- Optimizing $O$ uses a gradient descent method
  - Critically dependent on a good initialization

- Dimensionality of embedding (i.e., $d$) could be large
  - E.g., 8,353 for YouTube and 6,288,363 for Orkut [Yang & Leskovec 2012]
Sampling

- Objective function $O$ is a sum over $O(n^2)$ terms

$$O = \sum_{(i,j)\in E} \|X_i - X_j\|^2 + \alpha \cdot \sum_{(i,j)\notin E} \left(1 - \|X_i - X_j\|\right)^2, \alpha = \frac{m}{\binom{n}{2} - m}$$

- Observation: balancing factor $\alpha$ is close to 0
  - Very inefficient
  - Possible to approximately represent $O$ by sampling

- Sampled objective function $O$

$$O \approx \sum_{(i,j)\in E} \|X_i - X_j\|^2 + \sum_{(i,j)\in E_s} \left(1 - \|X_i - X_j\|\right)^2, E_s \subset \{(i,j) | (i,j) \notin E\}$$

- $|E_s| = |E| = m$
Graph partitioning based initialization

- Optimizing $O$ uses a gradient descent method
  - Critically dependent on a good initialization

- A good initialization means small value of $O$
  - Densely connected nodes have similar values of $X_i$
  - Nodes across groups have diverse values of $X_i$

- Incorporating graph partitioning (METIS) for initialization
  - $P_i$: partition number of node $i$

\[
X_i = (x_i^1, \ldots, x_i^d), \quad x_i^j = \begin{cases} 
1/\sqrt{2} & j = P_i \\
0 & j \neq P_i 
\end{cases}
\]
Dimension reduction

- Dimensionality of embedding (i.e., $d$) can be large

- The complete $d$-dimensions are unnecessary
  - Nodes typically connect to a limited number of communities
  - A limited number of communities suffice to ascertain anomalies

- Data approximation ($k+\beta$ reduction)
  - only maintain ($k+\beta$)-dimensions for embedding of each node
  - $k$ : the maximum number of communities to connect
  - $\beta$ : tolerate mistakes when determining the $k$ communities
  - $k << d$ & $\beta << d$, e.g., 10 & 2 for a network with $n = 10^6$
# Impacts of optimization techniques

<table>
<thead>
<tr>
<th></th>
<th>Space</th>
<th>Efficiency</th>
<th>Effectiveness</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Sampling</strong></td>
<td>/</td>
<td>Prev.: O($n^2 \cdot d$)</td>
<td>Remain effective (from experiments)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>After: O($m \cdot d$)</td>
<td></td>
</tr>
<tr>
<td><strong>Graph partitioning</strong></td>
<td>/</td>
<td>Prev.: 0</td>
<td>Provide a good initialization</td>
</tr>
<tr>
<td></td>
<td></td>
<td>After: O($n+m+d \cdot \log(d)$)</td>
<td></td>
</tr>
<tr>
<td><strong>k+β reduction</strong></td>
<td>Prev.: O($n \cdot d$)</td>
<td>Prev.: O($t \cdot m \cdot d$)</td>
<td>Slightly improve effectiveness</td>
</tr>
<tr>
<td></td>
<td>After: O($n \cdot (k+β)$)</td>
<td>$t$: # of iterations</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>After: O($t \cdot m \cdot (k+β)$)</td>
<td></td>
</tr>
</tbody>
</table>
Outline

- Anomaly detection model
- Algorithm optimizations
- Evaluation
Experimental settings

- **Datasets**

<table>
<thead>
<tr>
<th>Dataset</th>
<th># of nodes</th>
<th># of edges</th>
<th>Descriptions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon</td>
<td>334,863</td>
<td>925,872</td>
<td>Product co-purchasing</td>
</tr>
<tr>
<td>DBLP</td>
<td>1,150,852</td>
<td>5,098,175</td>
<td>Co-authorship</td>
</tr>
<tr>
<td>Synthetic</td>
<td>$10^5 - 4 \times 10^6$</td>
<td>$m = n^{1.15}$</td>
<td>LFR-benchmark graph</td>
</tr>
</tbody>
</table>

- Anomaly injection on Synthetic data for ground-truth of anomalies

- **Algorithms**

  - **Embed($d$)**: embedding of $d$-dimensions
  - **Embed($k+\beta$)**: embedding with $k+\beta$ reduction
  - **Oddball**: based on violation of power-laws of egonet-based features
  - **MDS($d$)**: similar to Embed($d$), except using multi-dimensional scaling for embedding (preserve global structure)

- **Parameters**: $d = n/500$, $k = \text{avgDeg}$, $\beta = k/4$

- **Implementation**: C++, Core i5 3.10GHz, 16GB of memory
Case study on DBLP

- Different people with the same name
  Wei Wang
  - 84 people named Wei Wang [DBLP, May 10 2016]
  - University of Waterloo (Canada), Fudan University (China), University of California, San Diego (USA), etc.

- People with many collaborators in diverse institutes
  Dr. Ajith Abraham
  - Director of intelligence research labs which has members from more than 100 countries
  - Work in a multi-disciplinary environment involving machine intelligence, cyber security, sensor networks and data mining
  - Teach in 23 universities all over the world
Quality study: modularity

- Modularity measures the strength of division of a network into communities.
- Using modularity to evaluate the improvement of the effectiveness of community detection.

<table>
<thead>
<tr>
<th></th>
<th>oddball</th>
<th>Embed(d)</th>
<th>Embed(k+β)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon</td>
<td>2.1%</td>
<td>2.8%</td>
<td>3.0%</td>
</tr>
<tr>
<td>DBLP</td>
<td>4.2%</td>
<td>4.1%</td>
<td>5.6%</td>
</tr>
</tbody>
</table>

Table 1: Improvement of modularity
Quality study: $F_1$ measure

- On Synthetic data with ground-truth of anomalies
- Mixing parameter $\mu$: fraction of inter-group edges (i.e., $\mu \uparrow$, strength of community structure $\downarrow$)

<table>
<thead>
<tr>
<th>Varying graph sizes</th>
<th>oddball</th>
<th>Embed($d$)</th>
<th>Embed($k+\beta$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Varying $\mu$</td>
<td>70%</td>
<td>88%</td>
<td>89%</td>
</tr>
<tr>
<td></td>
<td>68%</td>
<td>86%</td>
<td>88%</td>
</tr>
</tbody>
</table>

Table 2: $F_1$ score of anomalies
Impacts on quality: $d$ & embedding

- Synthetic data, $n = 400K$, $n/500 = 800$

<table>
<thead>
<tr>
<th></th>
<th>MDS($d$)</th>
<th>Embed($d$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$d = 200$</td>
<td>11.3%</td>
<td>89.4%</td>
</tr>
<tr>
<td>$d = 400$</td>
<td>13.6%</td>
<td>90.6%</td>
</tr>
<tr>
<td>$d = 600$</td>
<td>12.7%</td>
<td>89.8%</td>
</tr>
<tr>
<td>$d = 800$</td>
<td>7.9%</td>
<td>85.5%</td>
</tr>
<tr>
<td>$d = 1000$</td>
<td>11.3%</td>
<td>88.8%</td>
</tr>
<tr>
<td>Average</td>
<td>11.3%</td>
<td>88.8%</td>
</tr>
</tbody>
</table>

Table 3: MDS($d$) vs. Embed($d$) using $F_1$ measure

- Multi-dimensional scaling fails to effectively detect anomalies
- Our approach works well as long as $d$ falls into a reasonable range
Efficiency study

Table 4: running time comparison

<table>
<thead>
<tr>
<th></th>
<th>E(k+β)/E(d)</th>
<th>E(k+β)/MDS(d)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amazon</td>
<td>35.3%</td>
<td>25.0%</td>
</tr>
<tr>
<td>DBLP</td>
<td>23.4%</td>
<td>13.1%</td>
</tr>
<tr>
<td>Synthetic</td>
<td>25.6%</td>
<td>13.2%</td>
</tr>
</tbody>
</table>

x : out of memory exception
Summary

- Structural inconsistencies
  - Nodes that connect to a number of diverse influential communities
  - A formal quantitative definition of social brokers

- An embedding approach
  - Preserve local linkage structure of networks
  - A quantitative measure \textit{Ascore} inspired by structural inconsistencies and structural holes
  - Three algorithm optimization techniques

- Quality and efficiency results
  - Modularity increases 2.9%, 4.9% and 6.9% on Amazon, DBLP and Synthetic data
  - F1 measure is 88% on Synthetic data
  - Running time increases reasonably \textit{w.r.t} graph sizes
Thanks!

Q & A