Graph Mining: Applications

Karel Vaculík$^{1,2}$

$^1$KDLab, Faculty of Informatics Masaryk University, Brno
$^2$Gauss Algorithmic s.r.o., Brno

WIKT & Data a Znalosti 2016
Graph mining = data mining from graph (network) data

\[ G = (V, E) \]
Outline

1. Classification of Nodes
2. Anomaly Detection in Recommendation Networks
3. Anomaly Detection in Communication Networks
4. Community Detection in Voice-Call Networks
1. Classification of Nodes
1. Classification of Nodes

- Structural neighborhood-based classifier (SNBC)

- Assumes homophily and densely connected nodes to be correlated

- Structured random walk approach with variable termination probability

- Low- and high-degree nodes have lesser discriminative power
Datasets (multilabeled):

• **Youtube** – users in friendship relation; groups of interests as classes
• **PubMed** – citation network; three classes of publications
• **CoRA** – citation network; research topics
• **IMDb** – movies connected according to actor similarity; genres
• **Amazon Books** – similar books connected; genres
• **Wikipedia** (comp. science) – intralinks; 16 top-level categories
1. Classification of Nodes

SNBC results

10% of nodes for training

---

S. Nandanwar et al.: Structural Neighborhood Based Classification of Nodes in a Network. SIGKDD 2016.
2. Anomaly Detection in Recommendation Networks

2. Anomaly Detection in Recommendation Networks

- Supervised learning
- Semantic information from Google Knowledge Graph
- Structural (network) features from Related Places Graph

<table>
<thead>
<tr>
<th>t</th>
<th>Query</th>
<th>Entity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>taj mahal</td>
<td>taj mahal (landmark)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>trump taj mahal (casino)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>taj mahal (restaurant)</td>
</tr>
<tr>
<td>2</td>
<td>atlantic city casinos</td>
<td>borgota (casino)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>trump taj mahal (casino)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>trump plaza (casino)</td>
</tr>
<tr>
<td>3</td>
<td>trump plaza</td>
<td>trump plaza (casino)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>trump taj mahal (casino)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>borgota (casino)</td>
</tr>
</tbody>
</table>

(a) Hypothetical session queries
(b) Session-Entity Matrix
(c) Related Places Graph

Examples of anomalies detected by the system:

<table>
<thead>
<tr>
<th>Entity Location</th>
<th>Anomalous Recommendation(s) in Top-5</th>
<th>Reason</th>
</tr>
</thead>
<tbody>
<tr>
<td>IBM New York, NY (Software Company)</td>
<td>Victoria’s Secret (Intimate Apparel Store)</td>
<td>Nearby</td>
</tr>
<tr>
<td>Boys and Girls Club Orlando, FL (Youth Organization)</td>
<td>3 Strip Clubs (Adult Entertainment)</td>
<td>Auto-Completion &amp; Polysemy</td>
</tr>
<tr>
<td>Mom’s Bar Los Angeles, CA (Bar)</td>
<td>Los Angeles County Bar Association (Professional Association)</td>
<td>Ambiguous Phrases</td>
</tr>
<tr>
<td>Tony’s Small Engine Services Ashland, KY (Auto Repair Shop)</td>
<td>Academy Animal Hospital (Veternarian)</td>
<td>Data Sparsity</td>
</tr>
<tr>
<td>IKEA Canton, MI (Home Furnishings Store)</td>
<td>Comfort Suites (Hotel)</td>
<td>Unmodeled Phenomenon</td>
</tr>
<tr>
<td></td>
<td>Hampton Inn (Hotel)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>La Quinta Inn (Hotel)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Fairfield Inn (Hotel)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Extend Stay America (Hotel)</td>
<td></td>
</tr>
<tr>
<td>Florida Department of Agriculture Tampa, FL (State Government Agency)</td>
<td>Tampa Private Investigators (Private Detectives)</td>
<td>Conglomerate (issues firearm permits)</td>
</tr>
<tr>
<td></td>
<td>Equip 2 Conceal Firearms (Gun shop)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Shoot Straight (Gun shop)</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Florida Firearms Academy (Shooting Range)</td>
<td></td>
</tr>
</tbody>
</table>
3. Anomaly Detection in Communication Networks

- **ENRON** dataset
- employees sending emails
  - emails sent in one day => graph (snapshot)
- Series of snapshots (~ 900) = a *dynamic graph*
ENRON emails

DAY 678

DAY 679
3. Anomaly Detection in Communication Networks

**DGRMiner** (Dynamic Graph Rule Miner):

1. **Mine frequent graph patterns** (in the form of rules)

   ![Graph Pattern](image1)

   ![Graph Pattern](image2)

2. **Mine patterns deviating from the frequent ones**

   ![Graph Pattern](image3)

   ![Graph Pattern](image4)

---

Examples of anomalies:

Anomaly pattern:

Oulierness: 0.88

Explanation pattern:

Sup: 0.12, Conf: 0.61

Anomaly pattern:

Oulierness: 0.83

Explanation pattern:

Sup: 0.10, Conf: 0.71

3. Anomaly Detection in Communication Networks

Examples of anomalies:

Voice-call network

- Vertices = people (not only customers)
- Edges = aggregated calls
- Weights ≈ call traffic volume
  - Edges with low traffic filtered out
- ~ 1.8M vertices
- ~ 2.8M edges
4. Community Detection in Voice-Call Networks

Degree distribution
4. Community Detection in Voice-Call Networks

- Community detection by using weighted LPA
- ~ 400K communities found
- Distribution of community sizes:

One large community of size ~ 13,000 is omitted in the graph
Example communities of size 8
Sample subgraph with communities
Stability of the communities

- How fast communities change
  - after one week
  - after one month
  - ...
- Adjusted Rand index used for comparison
Utilization of detected communities:

- Churn prediction
- Detection of leaders in communities
- Identification of customers related to high-value customers
- Better offers for customers
- Households detection
Plenty of real-world graphs and associated tasks:

- classification (web pages; chemical compounds)
- clustering (community detection)
- link prediction (friendship suggestion)
- ranking (webpages)
- anomaly detection (fraud detection)
- pattern mining (graph DB indexing)
- ...