Anomaly detection in large graphs

Christos Faloutsos
CMU

www.cs.cmu.edu/~christos/TALKS/17-05-26-HKUST/
Thank you!

• Prof. Qiang Yang

• Prof. Lei CHEN

• Irene Lau
‘Hi’ to friends
Roadmap

• Introduction – Motivation
  – Why study (big) graphs?
• Part#1: Patterns in graphs
• Part#2: time-evolving graphs; tensors
• Conclusions
Graphs - why should we care?

>$10B; ~1B users
Graphs - why should we care?

Internet Map  [lumeta.com]

Food Web  [Martinez ’91]
Graphs - why should we care?

- web-log (‘blog’) news propagation
- computer network security: email/IP traffic and anomaly detection
- Recommendation systems
- ....

- Many-to-many db relationship -> graph
Motivating problems

• P1: patterns? Fraud detection?

• P2: patterns in time-evolving graphs / tensors
Motivating problems

• P1: patterns? Fraud detection?

• P2: patterns in time-evolving graphs / tensors
Motivating problems

• P1: patterns? Fraud detection?

• P2: patterns in time-evolving graphs / tensors

* Robust Random Cut Forest Based Anomaly Detection on Streams Sudipto Guha, Nina Mishra, Gourav Roy, Okke Schrijvers, ICML’16
Roadmap

• Introduction – Motivation
  – Why study (big) graphs?
• Part#1: Patterns & fraud detection
• Part#2: time-evolving graphs; tensors
• Conclusions
Part 1: Patterns, & fraud detection
Laws and patterns

• Q1: Are real graphs random?
Laws and patterns

• Q1: Are real graphs random?
• A1: NO!!
  – Diameter (‘6 degrees’; ‘Kevin Bacon’)
  – in- and out- degree distributions
  – other (surprising) patterns

• So, let’s look at the data
Solution# S.1

- Power law in the degree distribution [Faloutsos x 3 SIGCOMM99]

internet domains

log(degree) vs log(rank)

att.com

ibm.com
Solution# S.1

- Power law in the degree distribution [Faloutsos x 3 SIGCOMM99]

internet domains

log(degree) -0.82 log(rank)

att.com

ibm.com

HKUST, May'17
S2: connected component sizes

- Connected Components – 4 observations:

1.4B nodes
6B edges
S2: connected component sizes

- Connected Components

1) 10K x larger than next
**S2: connected component sizes**

- **Connected Components**

  2) ~0.7B singleton nodes
S2: connected component sizes

- Connected Components

3) SLOPE!
S2: connected component sizes

- Connected Components

4) Spikes!
S2: connected component sizes

- Connected Components

![Graph showing the distribution of connected component sizes with a focus on suspicious financial-advice sites (not existing now).]
Roadmap

• Introduction – Motivation
• Part#1: Patterns in graphs
  – P1.1: Patterns: Degree; Triangles
  – P1.2: Anomaly/fraud detection
• Part#2: time-evolving graphs; tensors
• Conclusions
Solution# S.3: Triangle ‘Laws’

• Real social networks have a lot of triangles
Solution# S.3: Triangle ‘Laws’

- Real social networks have a lot of triangles
  - Friends of friends are friends
- Any patterns?
  - 2x the friends, 2x the triangles?
Triangle Law: #S.3
[Tsourakakis ICDM 2008]

X-axis: degree
Y-axis: mean # triangles
n friends -> ~n^{1.6} triangles
Triangle counting for large graphs?

Anomalous nodes in Twitter (~ 3 billion edges)

[U Kang, Brendan Meeder, +, PAKDD’11]

HKUST, May (c) C. Faloutsos, 2017
Triangle counting for large graphs?

Anomalous nodes in Twitter (~ 3 billion edges)

[U Kang, Brendan Meeder, +, PAKDD’11]
Triangle counting for large graphs?

Anomalous nodes in Twitter (~3 billion edges)

[U Kang, Brendan Meeder, +, PAKDD’11]
Triangle counting for large graphs?

Anomalous nodes in Twitter (~ 3 billion edges)
[U Kang, Brendan Meeder, +, PAKDD ’11]

HKUST, May'17
(c) C. Faloutsos, 2017
Triangle counting for large graphs?

Anomalous nodes in Twitter (~ 3 billion edges)

[U Kang, Brendan Meeder, +, PAKDD’11]
# MORE Graph Patterns

<table>
<thead>
<tr>
<th>Static</th>
<th>Unweighted</th>
<th>Weighted</th>
</tr>
</thead>
<tbody>
<tr>
<td>✔️ L2. Triangle Power Law (TPL) [Tsourakakis `08]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L03. Eigenvalue Power Law (EPL) [Siganos et al. `03]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>L04. Community structure [Flake et al. <code>02, Girvan and Newman </code>02]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dynamic</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>✔️ L05. Densification Power Law (DPL) [Leskovec et al. `05]</td>
<td>✔️ L11. Weight Power Law (WPL) [McGlohon et al. `08]</td>
</tr>
<tr>
<td>L06. Small and shrinking diameter [Albert and Barabási <code>99, Leskovec et al. </code>05]</td>
<td></td>
</tr>
<tr>
<td>✔️ L07. Constant size 2nd and 3rd connected components [McGlohon et al. `08]</td>
<td></td>
</tr>
<tr>
<td>L08. Principal Eigenvalue Power Law ($\lambda_1$PL) [Akoglu et al. `08]</td>
<td></td>
</tr>
<tr>
<td>L09. Bursty/self-similar edge/weight additions [Gomez and Santonian <code>98, Gribble et al. </code>98, Crovella and</td>
<td></td>
</tr>
</tbody>
</table>

**RTG: A Recursive Realistic Graph Generator using Random Typing** Leman Akoglu and Christos Faloutsos. *PKDD’09.*
MORE Graph Patterns

<table>
<thead>
<tr>
<th>Unweighted</th>
<th>Weighted</th>
</tr>
</thead>
<tbody>
<tr>
<td>122. Triangle Power Law (TPL) [Tsourakakis '08]</td>
<td>139. Degree Power Law (DPL) [Leskovec et al. '05]</td>
</tr>
<tr>
<td>123. Eigenvalue Power Law (EPL) [Eigensatz et al. '03]</td>
<td>124. Small and shrinking diameter [Albert and Barabási '99, Leskovec et al. '05]</td>
</tr>
<tr>
<td>124. Community structure [Hale et al. '02, Girvan and Newman '02]</td>
<td>125. Constant size 2nd and 3rd connected components [McGlohon et al. '08]</td>
</tr>
</tbody>
</table>


Roadmap

• Introduction – Motivation
• Part#1: Patterns in graphs
  – P1.1: Patterns
  – P1.2: Anomaly / fraud detection
    • No labels – spectral
    • With labels: Belief Propagation
• Part#2: time-evolving graphs; tensors
• Conclusions
How to find ‘suspicious’ groups?

• ‘blocks’ are normal, right?
Except that:

• ‘blocks’ are normal, right?
• ‘hyperbolic’ communities are more realistic
  [Araujo+, PKDD’14]
Except that:

• ‘blocks’ are usually suspicious
• ‘hyperbolic’ communities are more realistic
  [Araujo+, PKDD’14]

Q: Can we spot blocks, easily?
Except that:

• ‘blocks’ are usually suspicious
• ‘hyperbolic’ communities are more realistic
  [Araujo+, PKDD’14]

Q: Can we spot blocks, easily?
A: Silver bullet: SVD!
Crush intro to SVD

• Recall: (SVD) matrix factorization: finds blocks

\[
\begin{align*}
\mathbf{U} & \approx \mathbf{X} \mathbf{V}^T \\
\mathbf{X} & = \mathbf{U} \mathbf{V}^T
\end{align*}
\]

where

- \( \mathbf{U} \) is the left singular matrix
- \( \mathbf{V} \) is the right singular matrix
- \( \mathbf{X} \) is the data matrix

\( N \) fans

\( M \) idols

\( \mathbf{u}_1 \)

\( \mathbf{v}_1 \)

‘music lovers’ ‘sports lovers’ ‘citizens’ ‘singers’ ‘athletes’ ‘politicians’

HKUST, May'17

(c) C. Faloutsos, 2017
Crush intro to SVD

• Recall: (SVD) matrix factorization: finds blocks

\[
\begin{align*}
&\mathbf{u}_1 \approx \mathbf{u}_1 + \mathbf{u}_i \\
&\mathbf{v}_1 \\
\end{align*}
\]
Inferring Strange Behavior from Connectivity Pattern in Social Networks

PAKDD’14

Meng Jiang, Peng Cui, Shiqiang Yang (Tsinghua)
Alex Beutel, Christos Faloutsos (CMU)
Lockstep and Spectral Subspace Plot

• Case #0: No lockstep behavior in random power law graph of 1M nodes, 3M edges
• Random ↔ “Scatter”

HKUST, May'17
**Lockstep and Spectral Subspace Plot**

- Case #1: non-overlapping lockstep
- “Blocks” ←→ “Rays”

**Adjacency Matrix**

**Spectral Subspace Plot**

---

Rule 1 (short “rays”): two blocks, high density (90%), no “camouflage”, no “fame”

HKUST, May'17

(c) C. Faloutsos, 2017
Lockstep and Spectral Subspace Plot

- Case #2: non-overlapping lockstep
- “Blocks; low density” ←→ Elongation

Adjacency Matrix  Spectral Subspace Plot

Rule 2 (long “rays”): two blocks, low density (50%), no “camouflage”, no “fame”

HKUST, May’17  (c) C. Faloutsos, 2017
Lockstep and Spectral Subspace Plot

• Case #3: non-overlapping lockstep
• “Camouflage” (or “Fame”) ←→ Tilting

“Rays”
Adjacency Matrix       Spectral Subspace Plot

Rule 3 (tilting “rays”): two blocks, with “camouflage”, no “fame”

HKUST, May'17
(c) C. Faloutsos, 2017
Lockstep and Spectral Subspace Plot

- Case #3: non-overlapping lockstep
- “Camouflage” (or “Fame”) ↔ Tilting “Rays”

Adjacency Matrix

Spectral Subspace Plot

Rule 3 (tilting “rays”): two blocks, no “camouflage”, with “fame”
Lockstep and Spectral Subspace Plot

- Case #4: ? lockstep
- “?” “Pearls”

Adjacency Matrix

Spectral Subspace Plot

HKUST, May'17
(c) C. Faloutsos, 2017
**Lockstep and Spectral Subspace Plot**

- Case #4: **overlapping** lockstep
- “Staircase” ←→ “Pearls”

---

**Adjacency Matrix**

**Spectral Subspace Plot**

Rule 4 ("pearls"): a “staircase” of three partially overlapping blocks.
Dataset

- Tencent Weibo
- 117 million nodes (with profile and UGC data)
- 3.33 billion directed edges
Real Data

- Spikes on the out-degree distribution
Roadmap

• Introduction – Motivation
• Part#1: Patterns in graphs
  – P1.1: Patterns
  – P1.2: Anomaly / fraud detection
    • No labels – spectral methods
    • With labels: Belief Propagation
• Part#2: time-evolving graphs; tensors
• Conclusions

HKUST, May'17
(c) C. Faloutsos, 2017
E-bay Fraud detection

w/ Polo Chau & Shashank Pandit, CMU
[www’07]
E-bay Fraud detection
E-bay Fraud detection
E-bay Fraud detection - NetProbe

HKUST, May'17

(c) C. Faloutsos, 2017
Popular press

And less desirable attention:
• E-mail from ‘Belgium police’ (‘copy of your code?’)
Summary of Part#1

• *many* patterns in real graphs
  – Power-laws everywhere
  – Long (and growing) list of tools for anomaly/fraud detection
Roadmap

• Introduction – Motivation
• Part#1: Patterns in graphs
• Part#2: time-evolving graphs
  – P2.1: tools/tensors
  – P2.2: other patterns
• Conclusions
Part 2: Time evolving graphs; tensors
Graphs over time -> tensors!

• Problem #2.1:
  – Given who calls whom, and when
  – Find patterns / anomalies
Graphs over time -> tensors!

• Problem #2.1:
  – Given who calls whom, and when
  – Find patterns / anomalies
Graphs over time -> tensors!

• Problem #2.1:
  – Given who calls whom, and when
  – Find patterns / anomalies
Graphs over time -> tensors!

• Problem #2.1:
  – Given who calls whom, and when
  – Find patterns / anomalies
Crush intro to SVD

• Recall: (SVD) matrix factorization: finds blocks

HKUST, May'17
(c) C. Faloutsos, 2017
Answer: tensor factorization

• PARAFAC decomposition

HKUST, May'17
(c) C. Faloutsos, 2017
Answer: tensor factorization

- PARAFAC decomposition
- Results for who-calls-whom-when
  - 4M x 15 days

HKUST, May'17
(c) C. Faloutsos, 2017
Anomaly detection in time-evolving graphs

- Anomalous communities in phone call data:
  - European country, 4M clients, data over 2 weeks

1 caller  5 receivers  4 days of activity

~200 calls to EACH receiver on EACH day!

HKUST, May'17  (c) C. Faloutsos, 2017
Anomaly detection in time-evolving graphs

- Anomalous communities in phone call data:
  - European country, 4M clients, data over 2 weeks

~200 calls to EACH receiver on EACH day!
Anomaly detection in time-evolving graphs

- Anomalous communities in phone call data:
  - European country, 4M clients, data over 2 weeks

Part 2: Conclusions

- Time-evolving / heterogeneous graphs -> tensors
- PARAFAC finds patterns
- Surprising temporal patterns (P.L. growth)
Roadmap

• Introduction – Motivation
  – Why study (big) graphs?
• Part#1: Patterns in graphs
• Part#2: time-evolving graphs; tensors
• Acknowledgements and Conclusions
Thanks

Disclaimer: All opinions are mine; not necessarily reflecting the opinions of the funding agencies

Thanks to: NSF IIS-0705359, IIS-0534205, CTA-INARC; Yahoo (M45), LLNL, IBM, SPRINT, Google, INTEL, HP, iLab
CONCLUSION#1 – Big data

- Patterns ≠ Anomalies

- Large datasets reveal patterns/outliers that are invisible otherwise
CONCLUSION#2 – tensors

• powerful tool

HKUST, May'17  (c) C. Faloutsos, 2017
References

• D. Chakrabarti, C. Faloutsos: *Graph Mining – Laws, Tools and Case Studies*, Morgan Claypool 2012
• http://www.morganclaypool.com/doi/abs/10.2200/S00449ED1V01Y201209DMK006
TAKE HOME MESSAGE:

Cross-disciplinarity
Thank you!

Cross-disciplinarity

www.cs.cmu.edu/~christos/TALKS/17-05-26-HKUST/