Mining Large Graphs: Patterns, Anomalies, and Fraud Detection

Christos Faloutsos

CMU
Thank you!

• Prof. Julian McAuley

• Nicholas Urioste
Roadmap

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

~1B nodes (web sites)
~6B edges (http links)
‘YahooWeb graph’
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

![Diagram showing log(rank) vs. log(degree) for internet domains with points marked for att.com and ibm.com.](image)

---

UCSD'16 (c) 2016, C. Faloutsos
Solution# S.1

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

internet domains

log(degree)

log(rank)

1000

exp(6.55063) \times x ^{-0.825116}

-0.82

att.com

ibm.com
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

<table>
<thead>
<tr>
<th>Size</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>10^9</td>
<td></td>
</tr>
<tr>
<td>10^8</td>
<td></td>
</tr>
<tr>
<td>10^7</td>
<td></td>
</tr>
<tr>
<td>10^6</td>
<td></td>
</tr>
<tr>
<td>10^5</td>
<td></td>
</tr>
<tr>
<td>10^4</td>
<td></td>
</tr>
<tr>
<td>10^3</td>
<td></td>
</tr>
<tr>
<td>10^2</td>
<td></td>
</tr>
<tr>
<td>10^1</td>
<td></td>
</tr>
<tr>
<td>10^0</td>
<td></td>
</tr>
</tbody>
</table>

suspicious financial-advice sites (not existing now)
Roadmap

• Introduction – Motivation
• Part#1: Patterns in graphs
  – Patterns: Degree; Triangles
  – Anomaly/fraud detection
  – Graph understanding
• 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 $\rightarrow \sim n^{1.6}$ triangles
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]

(c) 2016, C. Faloutsos
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]
# MORE Graph Patterns

<table>
<thead>
<tr>
<th>Unweighted</th>
<th>Weighted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Static</td>
<td></td>
</tr>
<tr>
<td>✔️ 2. Triangle Power Law (TPL) [Tsourakakis '08]</td>
<td></td>
</tr>
<tr>
<td>✔️ 3. Eigenvalue Power Law (EPL) [Siganos et al. '03]</td>
<td></td>
</tr>
<tr>
<td>✔️ 4. Community structure [Flake et al. '02, Girvan and Newman '02]</td>
<td></td>
</tr>
<tr>
<td>Dynamic</td>
<td></td>
</tr>
<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 '99, Leskovec et al. '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 Santonja '98, Gribble et al. '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>l02. Triangle Power Law (TPL) [Touroakis ’08]</td>
<td></td>
</tr>
<tr>
<td>l03. Eigenvalue Power Law (EPL) [Eiganos et al. ’03]</td>
<td></td>
</tr>
<tr>
<td>l04. Community structure [Blake et al. ’02, Girvan and Newman ’02]</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dynamic</th>
</tr>
</thead>
<tbody>
<tr>
<td>l05. Densification Power Law (DPL) [Leskovec et al. ’05]</td>
</tr>
<tr>
<td>l06. Small and shrinking diameter [Albert and Barabasi ’99, Leskovec et al. ’05]</td>
</tr>
<tr>
<td>l07. Constant size 2nd and 3rd connected components [McGlohon et al. ’08]</td>
</tr>
<tr>
<td>l08. Principal Eigenvalue Power Law (PPL) [Akoglu et al. ’08]</td>
</tr>
<tr>
<td>l11. Weight Power Law (WPL) [McGlohon et al. ’08]</td>
</tr>
</tbody>
</table>

• Mary McGlohon, Leman Akoglu, Christos Faloutsos. *Statistical Properties of Social Networks*. in "Social Network Data Analytics" (Ed.: Charu Aggarwal)

Roadmap

• Introduction – Motivation
• Part#1: Patterns in graphs
  – Patterns
  – 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

- M idols
- N fans

- 'music lovers' 'sports lovers'
- 'singers' 'athletes'
- 'citizens' 'politicians'

\[ \mathbf{U} \mathbf{V}^T \approx \mathbf{X} \]

(c) 2016, C. Faloutsos
Crush intro to SVD

• Recall: (SVD) matrix factorization: finds blocks

UCSD'16
(c) 2016, C. Faloutsos
Inferring Strange Behavior from Connectivity Pattern in Social Networks

Meng Jiang, Peng Cui, Shiqiang Yang
(Tsinghua, Beijing)

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”
**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”

UCSD'16  
(c) 2016, C. Faloutsos
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”

UCSD'16  (c) 2016, C. Faloutsos
**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”

UCSD'16  
(c) 2016, C. Faloutsos
**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”

(c) 2016, C. Faloutsos
Lockstep and Spectral Subspace Plot

• Case #4: lockstep
• “?” “Pearls”
Dataset

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

“Rays”

“Block”

“Pearls”

“Staircase”
Real Data

- Spikes on the out-degree distribution
Roadmap

• Introduction – Motivation

• Part#1: Patterns in graphs
  – Patterns
  – Anomaly / fraud detection
    • No labels – spectral methods
      – Suspiciousness
    • With labels: Belief Propagation

• Part#2: time-evolving graphs; tensors

• Conclusions
Suspicious Patterns in Event Data

2-modes

\[ ? \]

\[ ? \]

\[ ? \]

\[ ? \]

A General Suspiciousness Metric for Dense Blocks in Multimodal Data, Meng Jiang, Alex Beutel, Peng Cui, Bryan Hooi, Shiqiang Yang, and Christos Faloutsos, *ICDM*, 2015.
Suspicious Patterns in Event Data

Which is more suspicious?

20,000 Users
Retweeting same 20 tweets
6 times each
All in 10 hours

225 Users
Retweeting same 1 tweet
15 times each
All in 3 hours
All from 2 IP addresses

Answer: volume \* D_{KL}(p\|\|p_{background})
Suspicious Patterns in Event Data

Retweeting: “Galaxy Note Dream Project: Happy Happy Life Traveling the World”

<table>
<thead>
<tr>
<th>#</th>
<th>User × tweet × IP × minute</th>
<th>Mass</th>
<th>Suspiciousness</th>
</tr>
</thead>
<tbody>
<tr>
<td>CROSSSPOT</td>
<td>1 14 × 1 × 2 × 1,114</td>
<td>41,396</td>
<td>1,239,865</td>
</tr>
<tr>
<td></td>
<td>2 225 × 1 × 2 × 200</td>
<td>27,313</td>
<td>777,781</td>
</tr>
<tr>
<td></td>
<td>3 8 × 2 × 4 × 1,872</td>
<td>17,701</td>
<td>491,323</td>
</tr>
<tr>
<td>HOSVD</td>
<td>1 24 × 6 × 11 × 439</td>
<td>3,582</td>
<td>131,113</td>
</tr>
<tr>
<td></td>
<td>2 18 × 4 × 5 × 223</td>
<td>1,942</td>
<td>74,087</td>
</tr>
<tr>
<td></td>
<td>3 14 × 2 × 1 × 265</td>
<td>9,061</td>
<td>381,211</td>
</tr>
</tbody>
</table>
Roadmap

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

Suspicious fraudster -- this user has been behaving much like the other suspects by trading with the similar sets of possible accomplices.
Popular press

![Image of newspapers](usa_today.png), The Washington Post, Los Angeles Times

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

• Introduction – Motivation
• Part#1: Patterns in graphs
  – Patterns
  – Anomaly / fraud detection
    • No labels - Spectral methods
    • w/ labels: Belief Propagation – closed formulas
• Part#2: time-evolving graphs; tensors
• Conclusions
Unifying Guilt-by-Association Approaches: Theorems and Fast Algorithms

Danai Koutra*
U Kang
Hsing-Kuo Kenneth Pao

Tai-You Ke
Duen Horng (Polo) Chau
Christos Faloutsos

(*KDD dissertation award, 2016)

ECML PKDD, 5-9 September 2011, Athens, Greece
Problem Definition:
GBA techniques

Given: Graph; & few labeled nodes
Find: labels of rest (assuming network effects)
Are they related?

• RWR (Random Walk with Restarts)
  – google’s pageRank (‘if my friends are important, I’m important, too’)

• SSL (Semi-supervised learning)
  – minimize the differences among neighbors

• BP (Belief propagation)
  – send messages to neighbors, on what you believe about them
Are they related? YES!

- **RWR (Random Walk with Restarts)**
  - google’s pageRank (‘if my friends are important, I’m important, too’)

- **SSL (Semi-supervised learning)**
  - minimize the differences among neighbors

- **BP (Belief propagation)**
  - send messages to neighbors, on what you believe about them
## Correspondence of Methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Matrix</th>
<th>Unknown</th>
<th>known</th>
</tr>
</thead>
<tbody>
<tr>
<td>RWR</td>
<td>$[I - c \ A D^{-1}]$ × $x$ = $(1-c)y$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SSL</td>
<td>$[I + a(D - A)]$ × $x$ = $y$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>FABP</td>
<td>$[I + a D - c'A]$ × $b_h$ = $\phi_h$</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Diagram

- **Adjacency matrix**
  - $d_1$, $d_2$, $d_3$
  - $0$ $1$ $0$
  - $1$ $0$ $1$
  - $0$ $1$ $0$

- **Final labels/beliefs**
  - $0$
  - $1$
  - $1$

- **Prior labels/beliefs**
  - $0$
  - $1$

---

UCSD'16 (c) 2016, C. Faloutsos 67
Problem: e-commerce ratings fraud

- **Given** a heterogeneous graph on users, products, sellers and positive/negative ratings with “seed labels”
- **Find** the top $k$ most fraudulent users, products and sellers

Dhivya Eswaran, Stephan Günnemann, Christos Faloutsos, “ZooBP: Belief Propagation for Heterogeneous Networks”, In submission to VLDB 2017
ZooBP: Salient features

Fast approximate heterogeneous belief propagation with precise convergence guarantees!

Near-perfect accuracy  Scalable: linear in graph size

Dhivya Eswaran, Stephan Günnemann, Christos Faloutsos, “ZooBP: Belief Propagation for Heterogeneous Networks”, In submission to VLDB 2017
Summary of Part#1

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

Patterns \(\neq\) anomalies
Roadmap

• Introduction – Motivation
• Part#1: Patterns in graphs
• Part#2: time-evolving graphs; tensors
  – P2.1: time-evolving graphs
  – P2.2: inter-arrival time 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
Graphs over time -> tensors!

- Problem #2.1’:
  - Given author-keyword-date
  - Find patterns / anomalies

MANY more settings, with >2 ‘modes’
Graphs over time -> tensors!

• Problem #2.1’’:
  – Given subject – verb – object facts
  – Find patterns / anomalies

MANY more settings, with >2 ‘modes’
Graphs over time -> tensors!

- Problem #2.1’’’:
  - Given <triplets>
  - Find patterns / anomalies

MANY more settings, with >2 ‘modes’
(and 4, 5, etc modes)
Roadmap

• Introduction – Motivation
• Part#1: Patterns in graphs
• Part#2: time-evolving graphs; tensors
  – P2.1: time-evolving graphs
  – P2.2: inter-arrival time patterns
• Conclusions
Crush intro to SVD

• Recall: (SVD) matrix factorization: finds blocks
Answer to both: tensor factorization

- PARAFAC decomposition

$$\text{subject} \times \text{verb} \times \text{object} = \text{politicians} + \text{artists} + \text{athletes}$$
Answer: tensor factorization

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

Roadmap

• Introduction – Motivation
  – Why study (big) graphs?

• Part#1: Patterns in graphs

• Part#2: time-evolving graphs;
  – P2.1: tensors
  – P2.2: Inter-arrival time patterns

• Acknowledgements and Conclusions
KDD 2015 – Sydney, Australia

RSC: Mining and Modeling Temporal Activity in Social Media

Alceu F. Costa*    Yuto Yamaguchi    Agma J. M. Traina
Caetano Traina Jr.    Christos Faloutsos

*alceufc@icmc.usp.br
Pattern Mining: Datasets

<table>
<thead>
<tr>
<th>Reddit Dataset</th>
<th>Twitter Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time-stamp from comments</td>
<td>Time-stamp from tweets</td>
</tr>
<tr>
<td>21,198 users</td>
<td>6,790 users</td>
</tr>
<tr>
<td>20 Million time-stamps</td>
<td>16 Million time-stamps</td>
</tr>
</tbody>
</table>

For each user we have:

Sequence of postings time-stamps: \( T = (t_1, t_2, t_3, ...) \)

Inter-arrival times (IAT) of postings: \( (\Delta_1, \Delta_2, \Delta_3, ...) \)
Pattern Mining

**Pattern 1:** Distribution of IAT is heavy-tailed
Users can be inactive for long periods of time before making new postings

IAT Complementary Cumulative Distribution Function (CCDF)
(log-log axis)

Reddit Users

Twitter Users

(c) 2016, C. Faloutsos
Pattern Mining

**Pattern 1:** Distribution of IAT is heavy-tailed

Users can be inactive for long periods of time before making new postings

IAT Complementary Cumulative Distribution Function (CCDF)

No surprises – Should we give up?

Reddit Users

Twitter Users

UCSD'16

(c) 2016, C. Faloutsos
Human? Robots?

linear

log

PDF

0

0.005

0.01

10^2

10^3

10^4

10^5

Δ, IAT (seconds)

PDF

0

0.05

10^2

10^3

10^4

10^5

Δ, IAT (seconds)
Human? Robots?

2'  3h  1day

linear

log

PDF

Δ, IAT (seconds)

PDF

Δ, IAT (seconds)
Experiments: Can RSC-Spotter Detect Bots?

Precision vs. Sensitivity Curves

Good performance: curve close to the top

Twitter

- Precision > 94%
- Sensitivity > 70%

With strongly imbalanced datasets

# humans >> # bots

UCSD'16 (c) 2016, C. Faloutsos
Experiments: Can RSC-Spotter Detect Bots?

Precision vs. Sensitivity Curves

Good performance: curve close to the top

Precision > 96%
Sensitivity > 47%

With strongly imbalanced datasets
# humans >> # bots

Reddit

UCSD'16
(c) 2016, C. Faloutsos
Part 2: Conclusions

- Time-evolving / heterogeneous graphs -> tensors
- PARAFAC finds patterns
- (GigaTensor/HaTen2 -> fast & scalable)
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
Cast

Akoglu, Leman
Araujo, Miguel
Beutel, Alex
Chau, Polo
Eswaran, Dhivya
Hooi, Bryan
Kang, U
Koutra, Danai
Papalexakis, Vagelis
Shah, Neil
Shin, Kijung
Song, Hyun Ah

UCSD'16
(c) 2016, C. Faloutsos
CONCLUSION#1 – Big data

• Patterns vs. Anomalies

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

• powerful tool
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

http://www.cs.cmu.edu/~christos/TALKS/16-10-UCSD/