Must-read Material

- [Graph mining textbook] Deepayan Chakrabarti and Christos Faloutsos
  *Graph Mining: Laws, Tools and Case Studies*, Morgan Claypool, 2012
  – Part I (patterns)

Must-read Material (cont’d)


Must-read Material

- Jure Leskovec, Jon Kleinberg, Christos Faloutsos Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations, KDD 2005, Chicago, IL, USA
Main outline

- Introduction
- Indexing
- Mining
  - Graphs – patterns
  - Graphs – generators and tools
  - Association rules
  - …

Outline

- Introduction – Motivation
- Problem#1: Patterns in graphs
- Problem#2: Scalability
- Conclusions

Graphs - why should we care?

- IR: bi-partite graphs (doc-terms)
- web: hyper-text graph
- … and more:
Graphs - why should we care?

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

Outline

• Introduction – Motivation
  • Problem#1: Patterns in graphs
    – Static graphs
    – Weighted graphs
    – Time evolving graphs
  • Problem#2: Scalability
  • Conclusions

Problem #1 - network and graph mining

• What does the Internet look like?
• What does FaceBook look like?
• What is ‘normal’/‘abnormal’?
• which patterns/laws hold?

• What does the Internet look like?
• What does FaceBook look like?
• What is ‘normal’/‘abnormal’?
• which patterns/laws hold?
  – To spot anomalies (rarities), we have to discover patterns
Problem #1 - network and graph mining

- What does the Internet look like?
- What does FaceBook look like?
- What is ‘normal’/‘abnormal’?
- which patterns/laws hold?
  - To spot anomalies (rarities), we have to discover patterns
  - Large datasets reveal patterns/anomalies that may be invisible otherwise…

Are real graphs random?

- random (Erdos-Reyni) graph – 100 nodes, avg degree = 2
- before layout
- after layout
- No obvious patterns
  (generated with: pajek
  http://vlado.fmf.uni-lj.si/pub/networks/pajek/)

Graph mining

- Are real graphs random?

Laws and patterns

- Are real graphs random?
- A: NO!!
  - Diameter
  - in- and out- degree distributions
  - other (surprising) patterns

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

- Power law in the degree distribution [SIGCOMM99]

\[ \log(\text{rank}) \quad \log(\text{degree}) \]

internet domains

\[ \log(\text{degree}) \]

\[ \log(\text{rank}) \]

\[ \text{att.com} \]

\[ \text{ibm.com} \]

-0.82

Q: So what?

Solution# S.1

- Q: So what?
- A1: # of two-step-away pairs: \( O(d_{\text{max}}^2) \sim 10M^2 \)

```
internet domains
att.com
ibm.com
```

- \( \log(\text{rank}) - \log(\text{degree}) = -0.82 \)
- ~0.8PB -> a data center(!)

Solution# S.2: Eigen Exponent \( E \)

- A2: power law in the eigenvalues of the adjacency matrix

```
\[ \lambda x = A x \]
```

Exponent = slope

\( E = -0.48 \)

May 2001

- [Mihail, Papadimitriou '02]: slope is \( \frac{1}{3} \) of rank exponent
But:
How about graphs from other domains?

More power laws:
• web hit counts [w/ A. Montgomery]

And numerous more
• # of sexual contacts
• Income [Pareto] – ‘80-20 distribution’
• Duration of downloads [Bestavros+]
• Duration of UNIX jobs (‘mice and elephants’)
• Size of files of a user
• …
• ‘Black swans’
Outline

- Introduction – Motivation
- Problem#1: Patterns in graphs
  - Static graphs
    - degree, diameter, eigen,
    - Triangles
  - Weighted graphs
  - Time evolving graphs

Solution# S.3: Triangle ‘Laws’

- Real social networks have a lot of triangles
  - Friends of friends are friends
- Any patterns?

Triangle Law: #S.3
[Tsourakakis ICDM 2008]

- HEP-TH
- ASN
- Epinions

X-axis: # of participating triangles
Y: count (~ pdf)
Triangle Law: #S.3
[Tsourakakis ICDM 2008]

X-axis: # of participating triangles
Y: count (~ pdf)

Triangle Law: #S.4
[Tsourakakis ICDM 2008]

X-axis: degree
Y-axis: mean # triangles
n friends -> ~n^{1.6} triangles

Triangle Law: Computations
[Tsourakakis ICDM 2008]

But: triangles are expensive to compute
(3-way join; several approx. algos)
Q: Can we do that quickly?
A: Yes!
#triangles = 1/6 Sum ( \lambda_i^3 )
(and, because of skewness (S2),
we only need the top few eigenvalues!)
Triangle Law: Computations

[Tsourakakis ICDM 2008]

Wikipedia graph 2006-Nov-04

= 31M nodes = 37M edges

(1021x, 97.4%)
(1377x, 94.7%)
(1329x, 92.8%)

100x+ speed-up, >90% accuracy

Triangle counting for large graphs?

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

Any other ‘laws’?

Yes!
Any other ‘laws’?

Yes!
- Small diameter (~ constant!) –
  - six degrees of separation / ‘Kevin Bacon’
  - small worlds [Watts and Strogatz]

Any other ‘laws’?

- Bow-tie, for the web [Kumar+ ‘99]
- IN, SCC, OUT, ‘tendrils’
- disconnected components

Any other ‘laws’?

- power-laws in communities (bi-partite cores)
  [Kumar+, ‘99]
  
  Log(count)

  2:3 core
  (m:n core)

  Log(m)

Any other ‘laws’?

- “Jellyfish” for Internet [Tauro+ ‘01]
- core: ~clique
- ~5 concentric layers
- many 1-degree nodes
EigenSpokes


Useful for fraud detection!

• Eigenvectors of adjacency matrix
  ▪ equivalent to singular vectors
  (symmetric, undirected graph)

A = UΣUT

N

15-826 (c) 2014 C. Faloutsos
45

46

N

15-826 (c) 2014 C. Faloutsos
47

48
EigenSpokes

- Eigenvectors of adjacency matrix
  - equivalent to singular vectors
    (symmetric, undirected graph)

\[ A = U \Sigma U^T \]

EE plot:
- Scatter plot of scores of u1 vs u2
- One would expect
  - Many points @ origin
  - A few scattered ~randomly

\[ \begin{array}{c}
\text{1st Principal component} \\
\end{array} \]

\[ \begin{array}{c}
\text{2nd Principal component} \\
\end{array} \]
EigenSpokes - pervasiveness

- Present in mobile social graph
  - across time and space

- Patent citation graph

EigenSpokes - explanation

Near-cliques, or near-bipartite-cores, loosely connected
Near-cliques, or near-bipartite-cores, loosely connected

So what?
- Extract nodes with high scores
- high connectivity
- Good “communities”

Bipartite Communities!
- patents from same inventor(s)
  - `cut-and-paste’ bibliography!

Useful for fraud detection!

Outline
- Introduction – Motivation
- Problem#1: Patterns in graphs
  - Static graphs
    - degree, diameter, eigen,
    - Triangles
  - Weighted graphs
- Time evolving graphs
- Problem#2: Scalability
- Conclusions
Observations on weighted graphs?

- A: yes - even more ‘laws’!

M. McGlohon, L. Akoglu, and C. Faloutsos
*Weighted Graphs and Disconnected Components: Patterns and a Generator.*
*SIG-KDD 2008*

**Observation W.1: Fortification**

*Q: How do the weights of nodes relate to degree?*

More donors, more $?

- ‘Reagan’ $10
- ‘Clinton’ $7

More donors, even more $

- ‘Reagan’ $10
- ‘Clinton’ $7

Orgs-Candidates

- e.g. John Kerry, $10M received, from 1K donors

Weight: super-linear on in-degree

- exponent ‘iw’: $1.01 < iw < 1.26$
Outline

• Introduction – Motivation
• Problem#1: Patterns in graphs
  – Static graphs
  – Weighted graphs
  – Time evolving graphs
• Problem#2: Scalability
• Conclusions

Problem: Time evolution

• with Jure Leskovec (CMU -> Stanford)

• and Jon Kleinberg (Cornell – sabb. @ CMU)

T.1 Evolution of the Diameter

• Prior work on Power Law graphs hints at slowly growing diameter:
  – diameter $\sim O(\log N)$
  – diameter $\sim O(\log \log N)$
• What is happening in real data?

• Diameter shrinks over time
**T.1 Diameter – “Patents”**

- Patent citation network
- 25 years of data
- @1999
  - 2.9 M nodes
  - 16.5 M edges

**T.2 Temporal Evolution of the Graphs**

- N(t) … nodes at time t
- E(t) … edges at time t
- Suppose that
  \[ N(t+1) = 2 \times N(t) \]
- Q: what is your guess for
  \[ E(t+1) = \, ? \times E(t) \]
- A: over-doubled!
  - But obeying the "Densification Power Law"

**T.2 Densification – Patent Citations**

- Citations among patents granted
- @1999
  - 2.9 M nodes
  - 16.5 M edges
- Each year is a datapoint
Outline

• Introduction – Motivation
• Problem#1: Patterns in graphs
  – Static graphs
  – Weighted graphs
  – Time evolving graphs
• Problem#2: Scalability
• Conclusions

More on Time-evolving graphs

M. McGlohon, L. Akoglu, and C. Faloutsos
Weighted Graphs and Disconnected Components: Patterns and a Generator.
SIG-KDD 2008

[ Gelling Point ]

• Most real graphs display a gelling point
• After gelling point, they exhibit typical behavior. This is marked by a spike in diameter.

Observation T.3: NLCC behavior

Q: How do NLCC’s emerge and join with the GCC?

(‘‘NLCC’’ = non-largest conn. components)
– Do they continue to grow in size?
– or do they shrink?
– or stabilize?
Observation T.3: NLCC behavior

Q: How do NLCC’s emerge and join with the GCC?

(“NLCC” = non-largest conn. components)
– Do they continue to grow in size?
– or do they shrink?
– or stabilize?

Timing for Blogs

• with Mary McGlohon (CMU->Google)
• Jure Leskovec (CMU->Stanford)
• Natalie Glance (now at Google)
• Mat Hurst (now at MSR)
[SDM’07]
Post popularity drops-off – exponentially?

POWER LAW!

Exponent? -1.6

1. close to -1.5: Barabasi’s stack model
2. and like the zero-crossings of a random walk

T.5: duration of phone calls

*Trajectories of Call: Duration Distribution of Mobile Phone Users*

Pedro O. S. Vaz de Melo, Leman Akoglu, Christos Faloutsos, Antonio A. F. Loureiro

PKDD 2010

---

**Probably, power law (?)**

![Graph showing a power law distribution](image)

---

**No Power Law!**

![Graph showing a non-power law distribution](image)

---

**‘TLaC: Lazy Contractor’**

- The longer a task (phonecall) has taken,
- The even longer it will take

Odds ratio =

\[
\text{Casualties}(<x) : \text{Survivors}(\geq x) = \text{power law}
\]
Data Description

- Data from a private mobile operator of a large city
  - 4 months of data
  - 3.1 million users
  - more than 1 billion phone records
- Over 96% of ‘talkative’ users obeyed a TLAC distribution (‘talkative’: >30 calls)

Outline

- Introduction – Motivation
- Problem#1: Patterns in graphs
- Problem#2: Scalability - PEGASUS
- Conclusions

Scalability

- Yahoo: 5Pb of data [Fayyad, KDD’07]
- Problem: machine failures, on a daily basis
- How to parallelize data mining tasks, then?
- A: map/reduce – hadoop (open-source clone)
  http://hadoop.apache.org/
Outline – Algorithms & results

<table>
<thead>
<tr>
<th></th>
<th>Centralized</th>
<th>Hadoop/PEGASUS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degree Distr.</td>
<td>old</td>
<td>old</td>
</tr>
<tr>
<td>Pagerank</td>
<td>old</td>
<td>old</td>
</tr>
<tr>
<td>Diameter/ANF</td>
<td>old</td>
<td>HERE</td>
</tr>
<tr>
<td>Conn. Comp</td>
<td>old</td>
<td>HERE</td>
</tr>
<tr>
<td>Triangles</td>
<td>done</td>
<td>HERE</td>
</tr>
<tr>
<td>Visualization</td>
<td>started</td>
<td></td>
</tr>
</tbody>
</table>

HADI for diameter estimation

- Radius Plots for Mining Tera-byte Scale Graphs U Kang, Charalampos Tsourakakis, Ana Paula Appel, Christos Faloutsos, Jure Leskovec, SDM’10
- Naively: diameter needs $O(N^{**2})$ space and up to $O(N^{**3})$ time – prohibitive (N~1B)
- Our HADI: linear on E (~10B)
  - Near-linear scalability wrt # machines
  - Several optimizations -> 5x faster

YahooWeb graph (120Gb, 1.4B nodes, 6.6 B edges)
- Largest publicly available graph ever studied.
YahooWeb graph (120Gb, 1.4B nodes, 6.6 B edges)
• Largest publicly available graph ever studied.

- 7 degrees of separation (!)
• Diameter: shrunk

Q: Shape?

YahooWeb graph (120Gb, 1.4B nodes, 6.6 B edges)
• Effective diameter: surprisingly small.
• Multi-modality (?)!
Radius Plot of GCC of YahooWeb.

YahooWeb graph (120Gb, 1.4B nodes, 6.6 B edges)
- effective diameter: surprisingly small.
- Multi-modality: probably mixture of cores.

Conjecture:

YahooWeb graph (120Gb, 1.4B nodes, 6.6 B edges)
- effective diameter: surprisingly small.
- Multi-modality: probably mixture of cores.
Outline – Algorithms & results

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Centralized</th>
<th>Hadoop/PEGASUS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Degree Distr.</td>
<td>old</td>
<td>old</td>
</tr>
<tr>
<td>Pagerank</td>
<td>old</td>
<td>old</td>
</tr>
<tr>
<td>Diameter/ANF</td>
<td>old</td>
<td>HERE</td>
</tr>
<tr>
<td>Conn. Comp</td>
<td>old</td>
<td>HERE</td>
</tr>
<tr>
<td>Triangles</td>
<td>HERE</td>
<td></td>
</tr>
<tr>
<td>Visualization</td>
<td>started</td>
<td></td>
</tr>
</tbody>
</table>

Generalized Iterated Matrix Vector Multiplication (GIMV)

PEGASUS: A Peta-Scale Graph Mining System - Implementation and Observations.
U Kang, Charalampos E. Tsourakakis, and Christos Faloutsos.
(ICDM) 2009, Miami, Florida, USA.
Best Application Paper (runner-up).
Example: GIM-V At Work

- Connected Components – 4 observations:

1) 10K x larger than next

2) ~0.7B singleton nodes

3) SLOPE!
Example: GIM-V At Work

- Connected Components

<table>
<thead>
<tr>
<th>Size</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>300-size cmpt</td>
<td>500</td>
</tr>
<tr>
<td>1100-size cmpt</td>
<td>65</td>
</tr>
</tbody>
</table>

Why?

(c) 2014 C. Faloutsos

GIM-V At Work

- Connected Components over Time
- LinkedIn: 7.5M nodes and 58M edges

Stable tail slope after the gelling point

(c) 2014 C. Faloutsos

Outline

- Introduction – Motivation
- Problem#1: Patterns in graphs
- Problem#2: Scalability
- Conclusions
OVERALL CONCLUSIONS – low level:

• Several new **patterns** (fortification, shrinking diameter, triangle-laws, conn. components, etc)

• New **tools**:
  – anomaly detection (OddBall), belief propagation, immunization

• **Scalability**: PEGASUS / hadoop

OVERALL CONCLUSIONS – high level

• **BIG DATA**: Large datasets reveal patterns/outliers that are invisible otherwise

References

• Leman Akoglu, Christos Faloutsos: *RTG: A Recursive Realistic Graph Generator Using Random Typing*. ECML/PKDD (1) 2009: 13-28

References

• Jure Leskovec, Jon Kleinberg and Christos Faloutsos

  Graphs over Time: Densification Laws, Shrinking Diameters and Possible Explanations, KDD 2005

  (Best Research paper award).

• Jure Leskovec, Deepayan Chakrabarti, Jon M.

  Kleinberg, Christos Faloutsos: Realistic,

  Mathematically Tractable Graph Generation and

  Evolution, Using Kronecker Multiplication. PKDD

  2005: 133-145

• Jimeng Sun, Yingleian Xie, Hui Zhang, Christos

  Faloutsos. Less is More: Compact Matrix

  Decomposition for Large Sparse Graphs, SDM,

  Minneapolis, Minnesota, Apr 2007.

• Jimeng Sun, Spiros Papadimitriou, Philip S. Yu,

  and Christos Faloutsos, GraphScope: Parameter-

  free Mining of Large Time-evolving Graphs ACM

  SIGKDD Conference, San Jose, CA, August 2007

• Jimeng Sun, Dacheng Tao, Christos

  Faloutsos: Beyond streams and graphs: dynamic tensor analysis. KDD 2006:

  374-383

• Hanghang Tong, Christos Faloutsos, and

  Jia-Yu Pan, Fast Random Walk with Restart and Its Applications, ICDM 2006,

  Hong Kong.

• Hanghang Tong, Christos Faloutsos,

  Center-Piece Subgraphs: Problem

  Definition and Fast Solutions, KDD 2006,

  Philadelphia, PA
References


(Project info)

www.cs.cmu.edu/~pegasus

Thanks to: NSF IIS-0705359, IIS-0534205, CTA-INARC; Yahoo (M45), LLNL, IBM, SPRINT, Google, INTEL, HP, iLab