Part 1: Graph Mining – patterns

Christos Faloutsos
CMU

Our goal:
Open source system for mining huge graphs:

PEGASUS project (PEta GrApH mining System)

- [www.cs.cmu.edu/~pegasus](http://www.cs.cmu.edu/~pegasus)
- code and papers

References

- D. Chakrabarti, C. Faloutsos: *Graph Mining – Laws, Tools and Case Studies*, Morgan Claypool 2012

Outline

- Introduction – Motivation
- Part#1: Patterns in graphs
- Part#2: Tools (Ranking, proximity)
- Conclusions
Graphs - why should we care?

- network of companies & board-of-directors members
- ‘viral’ marketing
- web-log (‘blog’) news propagation
- computer network security: email/IP traffic and anomaly detection
- ....

Outline

- Introduction – Motivation
  - Patterns in graphs
    - Patterns in Static graphs
    - Patterns in Weighted graphs
    - Patterns in Time evolving graphs

Tepper, CMU, April 4
(c) C. Faloutsos, 2017
Network and graph mining

- How does the Internet look like?
- How 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…

Topology

How does the Internet look like? Any rules?

(Looks random – right?)
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

Laws – degree distributions

- Q: avg degree is ~2 - what is the most probable degree?

<table>
<thead>
<tr>
<th>degree</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>??</td>
</tr>
</tbody>
</table>

Laws – degree distributions

- Q: avg degree is ~2 - what is the most probable degree?

<table>
<thead>
<tr>
<th>degree</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>
Solution S1 . Power-law: outdegree $O$

The plot is linear in log-log scale [FFF’99]

$freq = degree (-2.15)$

Solution S1’

- Power law in the degree distribution [SIGCOMM99]

Solution# S.2: Eigen Exponent $E$

- A2: power law in the eigenvalues of the adjacency matrix [Mihail, Papadimitriou ’02]: slope is ½ of rank exponent

Solution# S.2: Eigen Exponent $E$

- Exponent = slope

E = -0.48

May 2001
But:
How about graphs from other domains?

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

epinions.com
- who-trusts-whom [Richardson + Domingos, KDD 2001]

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
- Patterns in graphs
  - Patterns in Static graphs
    - Degree
    - Triangles
    - …
  - Patterns in Weighted graphs
  - Patterns in Time evolving graphs
- Generators

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]

X-axis: # of Triangles a node participates in
Y-axis: count of such nodes
Triangle Law: #S.3
[Tsourakakis ICDM 2008]

HEP-TH
ASN
Epinions

X-axis: # of Triangles
a node participates in
Y-axis: count of such nodes

Epinions

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

Reuter
SN
Epinions

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

Outline

• Introduction – Motivation
• Patterns in graphs
  – Patterns in Static graphs
  – Patterns in Weighted graphs
  – Patterns in Time evolving graphs
• Generators

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 $?

More donors, even more $

Observation W.1: fortification: Snapshot Power Law

- Weight: super-linear on in-degree
- Exponent ‘iw’: $1.01 < iw < 1.26$

Outline

- Introduction – Motivation
- Patterns in graphs
  - Patterns in Static graphs
  - Patterns in Weighted graphs
  - Patterns in Time evolving graphs
- Generators

Orgs-Candidates

- e.g. John Kerry, $10M received, from 1K donors
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 ~ \(O(\log N)\)
  – diameter ~ \(O(\log \log N)\)

• What is happening in real data?

T.1 Diameter – “Patents”

• Patent citation network
• 25 years of data
• \(\text{at} 1999\)
  – 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
  \[ E(t+1) = ? \]
  \[ 2 \times E(t) \]

- 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
- Patterns in graphs
  - Patterns in Static graphs
  - Patterns in Weighted graphs
  - Patterns in Time evolving graphs
- Generators
More on Time-evolving graphs

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

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?

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

<table>
<thead>
<tr>
<th>Size</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>~0.7B</td>
<td>singleton nodes</td>
</tr>
</tbody>
</table>

Why?

300-size cmpt
X 500
1100-size cmpt
X 65

Why?
Example: GIM-V At Work

- Connected Components

Timing for Blogs

- with Mary McGlohon (CMU)
- Jure Leskovec (CMU->Stanford)
- Natalie Glance (now at Google)
- Mat Hurst (now at MSR)

T.4: popularity over time

Post popularity drops-off – exponentially?

Stable tail slope after the gelling point

Slope = 2.75

1 2 3

lag: days after post

@t

@t + lag
**T.4: Popularity over time**

Post popularity drops-off – exponentially? **POWER LAW!**
Exponent?

# in links \((\log)\)  
\(1\)  
\(2\)  
\(3\)  
\(\text{days after post (log)}\)

---

**Conclusions (part 1)**

**MANY** patterns in real graphs
- Skewed degree distributions
- Small (and shrinking) diameter
- Power-laws wrt triangles
- Oscillating size of connected components
- … and more

---

**References**

- D. Chakrabarti, C. Faloutsos: *Graph Mining – Laws, Tools and Case Studies*, Morgan Claypool 2012
References

- Jure Leskovec, Jon Kleinberg and Christos Faloutsos

Project info

www.cs.cmu.edu/~pegasus

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

Part 1
END