Large Graph Mining – Patterns, Tools and Cascade analysis

Christos Faloutsos
CMU
Roadmap

• Introduction – Motivation
  – Why ‘big data’
  – Why (big) graphs?
• Patterns in graphs
• Tools: fraud detection on e-bay
• Conclusions
Why ‘big data’

• Why big data?
• What is the problem definition?
• What are the major research challenges?
Main message:
Big data: often > experts

- ‘Super Crunchers’ *Why Thinking-By-Numbers is the New Way To Be Smart* by Ian Ayres, 2008

- Google won the machine translation competition 2005
Problem definition – big picture

Tera/Peta-byte data ➔ Analytics ➔ Insights, outliers

PSC-Sherlock, 2013

C. Faloutsos (CMU)
Roadmap

• Introduction – Motivation
  – Why ‘big data’
  – Why (big) graphs?
• Patterns in graphs
• Tools: fraud detection on e-bay
• Conclusions
Graphs - why should we care?

- $>$10B revenue
- $>$0.5B users
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
• ....
• Subject-verb-object -> graph
• Many-to-many db relationship -> graph
Outline

• Introduction – Motivation

• Patterns in graphs
  – Static graphs
  – Time evolving graphs
  – Radius, conn components

• Tools: fraud detection on e-bay

• Conclusions

PSC-Sherlock, 2013

C. Faloutsos (CMU)
Problem - network and graph mining

- What does the Internet look like?
- What does FaceBook look like?
- What is ‘normal’/‘abnormal’?
- which patterns/laws hold?
Problem - 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
Problem - 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…
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]

internet domains

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

\[
\text{att.com}
\]

\[
\text{ibm.com}
\]
Solution# S.1

- Power law in the degree distribution [SIGCOMM99]

internet domains

log(degree)  \[ \text{att.com} \]

log(rank)  -0.82  \[ \text{ibm.com} \]
But:

How about graphs from other domains?
More power laws:

- web hit counts [w/ A. Montgomery]

Web Site Traffic

Count (log scale)

Zipf

in-degree (log scale)

PSC-Sherlock, 2013

C. Faloutsos (CMU)
epinions.com

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

(count) degree

trusts-2000-people user

PSC-Sherlock, 2013

C. Faloutsos (CMU)
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’
Roadmap

• Introduction – Motivation
• Problem#1: Patterns in graphs
  – Static graphs
    • degree, diameter,
    • triangles
    • Cliques
  – Time evolving graphs
• …
Solution# S.2: Triangle ‘Laws’

- Real social networks have a lot of triangles
Solution# S.2: Triangle ‘Laws’

• Real social networks have a lot of triangles
  – Friends of friends are friends
• Any patterns?
Triangle Law: #S.2
[Tsourakakis ICDM 2008]

- Reuters
- SN
- Epinions

**X-axis:** degree
**Y-axis:** mean # triangles

$n$ friends $\rightarrow \sim$ triangles

PSC-Sherlock, 2013

CMU SCS
C. Faloutsos (CMU)
Triangle Law: #S.2
[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]
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]
Roadmap

• Introduction – Motivation

• Patterns in graphs
  – Static graphs
  – Time evolving graphs

• Tools

• ...

PSC-Sherlock, 2013
C. Faloutsos (CMU)
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?
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?

• 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 $E(t+1) =? 2 \times E(t)$
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

\[ N(t) \]

\[ E(t) \]

\[ E = 0.0002 \times 1.66 \]

\[ R^2 = 0.99 \]

PSC-Sherlock, 2013

C. Faloutsos (CMU)
T.3: popularity over time

Post popularity drops-off – exponentially?

lag: days after post
T.3: popularity over time

# in links (log)

Post popularity drops-off – exponentially?
POWER LAW!
Exponent?
T.3 : popularity over time

Post popularity drops-off – exponentially?
POWER LAW!
Exponent? -1.6
• close to -1.5: Barabasi’s stack model
• and like the zero-crossings of a random walk

PSC-Sherlock, 2013
C. Faloutsos (CMU)

-1.5 slope

\[
\text{Prob}(\text{RT} > x) \\
(\log)
\]

Response time (log)
Roadmap

• Introduction – Motivation
• Patterns in graphs
  – …
  – Diameter
  – Connected components
• Tools: fraud detection on e-bay
• Conclusions
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 ($N\sim1B$)

- Our HADI: linear on $E$ ($\sim10B$)
  - Near-linear scalability wrt # machines
  - Several optimizations -> 5x faster
19+ [Barabasi+]

~1999, ~1M nodes
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.
YahooWeb graph (120Gb, 1.4B nodes, 6.6 B edges)
• 7 degrees of separation (!)
• Diameter: shrunk
YahooWeb graph  (120Gb, 1.4B nodes, 6.6 B edges)
Q: Shape?

\[ \text{Radius} \]

\[ \text{Count} \]

\[ \text{Number of Nodes} \]

\[ \sim 7 \ (\text{undir.}) \]
YahooWeb graph (120Gb, 1.4B nodes, 6.6 B edges)
- effective diameter: surprisingly small.
- Multi-modality (?)!
Roadmap

• Introduction – Motivation
• Patterns in graphs
  – …
  – Diameter
  – Connected components
• Tools: fraud detection on e-bay
• Conclusions
Generalized Iterated Matrix Vector Multiplication (GIMV)

Example: GIM-V At Work

• Connected Components – 4 observations:
Example: GIM-V At Work

- Connected Components

![Diagram showing the distribution of connected components in Yahoo Web dataset with a 10Kx larger size than the next component.](image)
Example: GIM-V At Work

- Connected Components

Count

2) ~0.7B singleton nodes
Example: GIM-V At Work

- Connected Components

3) SLOPE!
Example: GIM-V At Work

• Connected Components

4) Spikes!
Example: GIM-V At Work

• Connected Components
Roadmap

• Introduction – Motivation
• Patterns in graphs
• Tools: fraud detection on e-bay
• 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
Popular press

And less desirable attention:

• E-mail from ‘Belgium police’ (‘copy of your code?’)
OVERALL CONCLUSIONS – low level:

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

• **New tools:**
  – belief propagation, gigaTensor, etc

• **Scalability:** PEGASUS / hadoop
OVERALL CONCLUSIONS – high level

- **BIG DATA**: Large datasets reveal patterns/outliers that are invisible otherwise
Thanks to: NSF IIS-0705359, IIS-0534205, CTA-INARC; Yahoo (M45), LLNL, IBM, SPRINT, Google, INTEL, HP, iLab

www.cs.cmu.edu/~pegasus
Cast

Akoglu, Leman
Beutel, Alex
Chau, Polo
Kang, U
Koutra, Danai

McGlohon, Mary
Prakash, Aditya
Papalexakis, Vagelis
Tong, Hanghang

PSC-Sherlock, 2013
C. Faloutsos (CMU)
Take-home message

Tera/Peta-byte data \[\Rightarrow\] Analytics \[\Rightarrow\] Insights, outliers

Big data reveal **insights** that would be invisible otherwise (even to **experts**)

PSC-Sherlock, 2013

C. Faloutsos (CMU)