Mining Large Graphs and Tensors - Patterns, Tools and Discoveries.

Christos Faloutsos
CMU
Thank you!

• Nikos Sidiropoulos

• Kuo-Chu Chang

• Zhi (Gerry) Tian
Roadmap

- Introduction – Motivation
  - Why ‘big data’
  - Why (big) graphs?
- Problem#1: Patterns in graphs
- Problem#2: Tools
- Conclusions
Why ‘big data’

• Why?
• What is the problem definition?
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

NSF, 3/2013
C. Faloutsos (CMU)
Problem definition – big picture

Tera/Peta-byte data

Analytics

Insights, outliers

Main emphasis in this talk

NSF, 3/2013

C. Faloutsos (CMU)
Roadmap

• Introduction – Motivation
  – Why ‘big data’
  – Why (big) graphs?
• Problem#1: Patterns in graphs
• Problem#2: Tools
• Problem#3: Scalability
• Conclusions
Graphs - why should we care?

>$10B revenue

>0.5B users

Food Web
[Martinez ’91]

Internet Map
[lumeta.com]
Graphs - why should we care?

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

- web-log (‘blog’) news propagation
- computer network security: email/IP traffic and anomaly detection
- ‘viral’ marketing
- Supplier-supply business chains (-> instabilities)
- ....
- Subject-verb-object -> graph
- Many-to-many db relationship -> graph
Outline

• Introduction – Motivation

• Problem#1: Patterns in graphs
  – Static graphs
  – Time evolving graphs

• Problem#2: Tools

• 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?
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

NSF, 3/2013

C. Faloutsos (CMU)
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…
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(degree)

LOG(ranking)

att.com

ibm.com
Solution# S.1

- Power law in the degree distribution
  [SIGCOMM99]

  internet domains

  - att.com
  - ibm.com

  log(rank)

  log(degree)

  -0.82
Solution# S.1

- Q: So what?

internet domains

log(degree)

log(rank)

att.com

ibm.com

-0.82
Solution# S.1

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

internet domains

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

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

\(~0.8\text{PB} \rightarrow \text{a data center(!)}\)
Solution# S.1

- Q: So what?
- A1: \# of two-step-away intersect

Such patterns -> New algorithms

\(~0.8PB -> a data center(!)\)

-0.82

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

\(\text{?)} \sim 10M^2\)
Solution# S.2: Eigen Exponent $E$

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

Exponent = slope

$E = -0.48$

May 2001
Many more power laws

• # 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, eigen,
    • triangles
    • cliques
  – Weighted graphs
  – Time evolving graphs

• Problem#2: Tools
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?
Triangle Law: #S.3
[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) – $O(d_{\text{max}}^2)$

Q: Can we do that quickly?

A:
Triangle Law: Computations
[Tsourakakis ICDM 2008]

But: triangles are expensive to compute
(3-way join; several approx. algos) – $O(d_{\text{max}}^2)$

Q: Can we do that quickly?
A: Yes!

$\#\text{triangles} = \frac{1}{6} \text{Sum} (\lambda_i^3)$
(and, because of skewness (S2),
we only need the top few eigenvalues! - $O(E)$)
Triangle Law: Computations

[Tsourakakis ICDM 2008]

Wikipedia graph 2006-Nov-04
≈ 3,1M nodes ≈ 37M edges

1000x+ speed-up, >90% accuracy
Triangle counting for large graphs?

Anomalous nodes in Twitter (~ 3 billion edges)

[U Kang, Brendan Meeder, +, PAKDD’11]

NSF, 3/2013  C. Faloutsos (CMU)
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
• Problem#1: Patterns in graphs
  – Static graphs
  – Time evolving graphs
• Problem#2: Tools
• ...
T.1: popularity over time

Post popularity drops-off – exponentially?

lag: days after post

# in links
T.1 : popularity over time

Post popularity drops-off – exponentially?

POWER LAW!

Exponent?
T.1 : 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

NSF, 3/2013
C. Faloutsos (CMU)
-1.5 slope


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

![Graph showing a line with -1.5 slope between response time (log) and probability (log)].
-1.5 slope

Roadmap

• Introduction – Motivation
• Problem#1: Patterns in graphs
• Problem#2: Tools
  – (Belief Propagation)
  – Tensors
  – Spike analysis
• Conclusions
GigaTensor: Scaling Tensor Analysis Up By 100 Times – Algorithms and Discoveries

U Kang
Evangelos Papalexakis
Abhay Harpale
Christos Faloutsos

KDD’12

NSF, 3/2013
Background: Tensor

• Tensors (=multi-dimensional arrays) are everywhere
  – Hyperlinks & anchor text [Kolda+,05]
Time evolving graphs: Tensors

\[ \chi \]

date

caller

callee

1
1
1
1
1
1
1
1
1
1
Background: Tensor

- Tensors (=multi-dimensional arrays) are everywhere
  - Sensor stream (time, location, type)
  - Predicates (subject, verb, object) in knowledge base

"Eric Clapton plays guitar"
"Barrack Obama is the president of U.S."

\[
\begin{array}{c}
\text{(48M) verbs} \\
\text{(26M) subjects} \\
\text{subjects} \\
\text{objects (26M)}
\end{array}
\]

NELL (Never Ending Language Learner) data
Nonzeros = 144M

NSF, 3/2013
C. Faloutsos (CMU)
Background: Tensor

- Tensors (=multi-dimensional arrays) are everywhere
  - Sensor stream (time, location, type)
  - Predicates (subject, verb, object) in knowledge base

Anomaly Detection in Computer networks
all I learned on tensors: from

Nikos Sidiropoulos
UMN

Tamara Kolda,
Sandia Labs
(tensor toolbox)

NSF, 3/2013
Problem Definition

• How to decompose a billion-scale tensor?
  – Corresponds to SVD in 2D case
Problem Definition

• How to decompose a billion-scale tensor?
  – Corresponds to SVD in 2D case = soft clustering
Problem Definition

- Q1: Dominant concepts/topics?
- Q2: Find synonyms to a given noun phrase?
- (and how to scale up: |data| > RAM)

NELL (Never Ending Language Learner) data
Nonzeros = 144M

(48M) verbs

(26M)

objects (26M)

subjects

x

NSF, 3/2013

C. Faloutsos (CMU)
Experiments

- **GigaTensor** solves \textit{100x} larger problem

\[
\text{Number of nonzero} = \frac{I}{50}
\]

\text{(I)} \quad \text{(J)} \quad \text{(K)}

\text{NSF, 3/2013}

C. Faloutsos (CMU)
A1: Concept Discovery

- Concept Discovery in Knowledge Base

<table>
<thead>
<tr>
<th>Concept 1: &quot;Web Protocol&quot;</th>
<th>Concept 2: &quot;Credit Cards&quot;</th>
<th>Concept 3: &quot;Health System&quot;</th>
<th>Concept 4: &quot;Family Life&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>internet protocol</td>
<td>credit information</td>
<td>health provider</td>
<td>life rest</td>
</tr>
<tr>
<td>file</td>
<td>debt</td>
<td>child providers</td>
<td>family part</td>
</tr>
<tr>
<td>data</td>
<td>number</td>
<td>home system</td>
<td>body years</td>
</tr>
<tr>
<td>software suite</td>
<td>'np1’ ‘card’ ‘np2’</td>
<td>'np1’ ‘care’ ‘np2’</td>
<td>'np2’ ‘of’ ‘my’ ‘np1’</td>
</tr>
<tr>
<td>‘np1’ ‘stream’ ‘np2’</td>
<td>'np1’ ‘report’ ‘np2’</td>
<td>'np1’ ‘service’ ‘np2’</td>
<td>'np2’ ‘of’ ‘his’ ‘np1’</td>
</tr>
<tr>
<td>‘np1’ ‘marketing’ ‘np2’</td>
<td>'np1’ ‘cards’ ‘np2’</td>
<td></td>
<td>'np2’ ‘of’ ‘her’ ‘np1’</td>
</tr>
</tbody>
</table>

NSF, 3/2013
C. Faloutsos (CMU)
# A1: Concept Discovery

<table>
<thead>
<tr>
<th>Concept</th>
<th>Noun Phrase 1</th>
<th>Noun Phrase 2</th>
<th>Context</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concept 1: &quot;Web Protocol&quot;</td>
<td>internet</td>
<td>protocol</td>
<td>‘np1’ ‘stream’ ‘np2’</td>
</tr>
<tr>
<td></td>
<td>file</td>
<td>software</td>
<td>‘np1’ ‘marketing’ ‘np2’</td>
</tr>
<tr>
<td></td>
<td>data</td>
<td>suite</td>
<td>‘np1’ ‘dating’ ‘np2’</td>
</tr>
<tr>
<td>Concept 2: &quot;Credit Cards&quot;</td>
<td>credit</td>
<td>information</td>
<td>‘np1’ ‘card’ ‘np2’</td>
</tr>
<tr>
<td></td>
<td>Credit</td>
<td>debt</td>
<td>‘np1’ ‘report’ ‘np2’</td>
</tr>
<tr>
<td></td>
<td>library</td>
<td>number</td>
<td>‘np1’ ‘cards’ ‘np2’</td>
</tr>
<tr>
<td>Concept 3: &quot;Health System&quot;</td>
<td>health</td>
<td>provider</td>
<td>‘np1’ ‘care’ ‘np2’</td>
</tr>
<tr>
<td></td>
<td>child</td>
<td>providers</td>
<td>‘np’ ‘insurance’ ‘np2’</td>
</tr>
<tr>
<td></td>
<td>home</td>
<td>system</td>
<td>‘np1’ ‘service’ ‘np2’</td>
</tr>
</tbody>
</table>
### A2: Synonym Discovery

<table>
<thead>
<tr>
<th>(Given) Noun Phrase</th>
<th>(Discovered) Potential Synonyms</th>
</tr>
</thead>
<tbody>
<tr>
<td>pollutants</td>
<td>dioxin, sulfur dioxide, greenhouse gases, particulates, nitrogen oxide, air pollutants, cholesterol</td>
</tr>
<tr>
<td>disabilities</td>
<td>infections, dizziness, injuries, diseases, drowsiness, stiffness, injuries</td>
</tr>
<tr>
<td>vodafone</td>
<td>verizon, comcast</td>
</tr>
<tr>
<td>Christian history</td>
<td>European history, American history, Islamic history, history</td>
</tr>
<tr>
<td>disbelief</td>
<td>dismay, disgust, astonishment</td>
</tr>
</tbody>
</table>
Roadmap

• Introduction – Motivation
• Problem#1: Patterns in graphs
• Problem#2: Tools
  – Belief propagation
  – Tensors
  – Spike analysis
  – Graph summarization
• Conclusions
Rise and fall patterns in social media

- **Meme (# of mentions in blogs)**
  - short phrases Sourced from U.S. politics in 2008
  - “you can put lipstick on a pig”

- “yes we can”
Rise and fall patterns in social media

- Can we find a unifying model, which includes these patterns?
  - **four** classes on YouTube [Crane et al. ’08]
  - **six** classes on Meme [Yang et al. ’11]
Rise and fall patterns in social media

• Answer: YES!

- We can represent all patterns by single model

In Matsubara+ SIGKDD 2012
Main idea - SpikeM

- 1. **Un-informed bloggers** (uninformed about rumor)
- 2. **External shock at time** $n_b$ (e.g., breaking news)
- 3. **Infection** (word-of-mouth)

Infectiveness of a blog-post at age $n$:

- $\beta$ – Strength of infection (quality of news)
- $f(n)$ – Decay function
Main idea - SpikeM

- 1. Un-informed bloggers (uninformed about rumor)
- 2. External shock at time $n_b$ (e.g., breaking news)
- 3. Infection (word-of-mouth)

Infectiveness of a blog-post at age $n$:

\[ \beta \] – Strength of infection (quality of news)

\[ f(n) \] – Decay function \[ f(n) = \beta \times n^{-1.5} \]
SpikeM - with periodicity

- Full equation of SpikeM

$$\Delta B(n+1) = p(n+1) \cdot \left[ U(n) \cdot \sum_{t=n_b}^{n} (\Delta B(t) + S(t)) \cdot f(n+1-t) + \epsilon \right]$$

Bloggers change their activity over time (e.g., daily, weekly, yearly)

Periodicity

noon

Peak

3am

Dip

activity

Time n

p(n)

NSF, 3/2013

C. Faloutsos (CMU)
Details

- Analysis – exponential rise and power-law fall

Lin-log

Log-log

Rise-part

SI $\rightarrow$ exponential
SpikeM $\rightarrow$ exponential

Carnegie Mellon

NSF, 3/2013
C. Faloutsos (CMU)
Details

- Analysis – exponential rise and power-law fall

\[ \text{SI} \rightarrow \text{exponential} \]
\[ \text{SpikeM} \rightarrow \text{power law} \]
Tail-part forecasts

- **SpikeM** can capture tail part
“What-if” forecasting

(1) First spike
(2) Release date
(3) Two weeks before release

e.g., given
(1) first spike,
(2) release date of two sequel movies
(3) access volume before the release date
“What-if” forecasting

(1) First spike
(2) Release date
(3) Two weeks before release

November 19, 2010
"Deathly Hallows part 1"

July 15, 2009
"Harry Potter and the Half-Blood Prince"

July 15, 2011
"Deathly Hallows part 2"

SpikeM can forecast upcoming spikes

NSF, 3/2013
C. Faloutsos (CMU)
Roadmap

• Introduction – Motivation
• Problem#1: Patterns in graphs
• Problem#2: Tools
  – Belief Propagation
  – Tensors
  – Spike analysis
  – Graph understanding (through MDL)
• Conclusions
Summarizing Graphs

Goal:

Main Idea: MDL + ‘syllables’:
star, clique, chain, bi-partite core

Koutra, Kang, Vreeken, et al, (subm.)
Summarizing Wiki-controversy

**Top-8 stars:** admins, bots

**Top-1 and top-2 bipartite cores:** edit wars.

**Left:** warring factions ('Kiev' vs 'Kyev')

**Right:** between vandals
Roadmap

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

• Several new patterns (power laws, triangle-laws, 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
(Graph) Analytics

Theory & Algo.

Biology

Physics

Comp. Systems

ML, Stats., DSP

Social Science

Econ.

NSF, 3/2013

C. Faloutsos (CMU)
Cross-disciplinarity: A must
References

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

References

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


References

• Christos Faloutsos, Tamara G. Kolda, Jimeng Sun: *Mining large graphs and streams using matrix and tensor tools*. Tutorial, SIGMOD Conference 2007: 1174
References

References

• Jure Leskovec, Jon Kleinberg and Christos Faloutsos

References

• Yasuko Matsubara, Yasushi Sakurai, B. Aditya Prakash, Lei Li, Christos Faloutsos, "Rise and Fall Patterns of Information Diffusion: Model and Implications", KDD’12, pp. 6-14, Beijing, China, August 2012
References


• 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
References

• Jimeng Sun, Dacheng Tao, Christos Faloutsos: *Beyond streams and graphs: dynamic tensor analysis*. KDD 2006: 374-383
References


• Hanghang Tong, Christos Faloutsos, *Center-Piece Subgraphs: Problem Definition and Fast Solutions*, KDD 2006, Philadelphia, PA
References

• Hanghang Tong, Christos Faloutsos, Brian Gallagher, Tina Eliassi-Rad: *Fast best-effort pattern matching in large attributed graphs.* KDD 2007: 737-746

References

Project info & ‘thanks’

www.cs.cmu.edu/~pegasus

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

Akoglu, Leman

Beutel, Alex

Chau, Polo

Kang, U

Koutra, Danai

McGlohon, Mary

Prakash, Aditya

Papalexakis, Vagelis

Tong, Hanghang

NSF, 3/2013

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**)

NSF, 3/2013

C. Faloutsos (CMU)