Large Graph Mining - Patterns, Explanations and Cascade Analysis

Christos Faloutsos
CMU
Thank you!

• Prof. Xiang Zhang

• Profs Meral & Tekin Ozsoyoglu

• Prof. Mike Lewicki
Roadmap

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

>$10B revenue

>0.5B users

Internet Map
[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
Roadmap

• Introduction – Motivation

• Part#1: Patterns in graphs
  – Static graphs
  – Time-evolving graphs
  – Why so many power-laws?

• Part#2: Cascade analysis

• Conclusions
Laws and patterns

• Q1: Are real graphs random?
Laws and patterns

• Q1: Are real graphs random?
  • A1: NO!!
    – Diameter
    – in- and out- degree distributions
    – other (surprising) patterns
• Q2: why ‘no good cuts’?
• A2: <self-similarity – stay tuned>

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

- Power law in the degree distribution
  [SIGCOMM99]

internet domains

log(degree) vs. log(rank)

att.com

ibm.com
Solution# S.1

- Power law in the degree distribution
  [SIGCOMM99]

internet domains

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

-0.82

att.com

ibm.com
Solution# S.1

- Q: So what?

![Graph showing internet domains with logarithmic axes for degree and rank. The graph includes points for att.com and ibm.com, with a slope of -0.82.]

(c) 2014, C. Faloutsos
Solution# S.1

• Q: So what?
  = friends of friends (F.O.F.)

• A1: # of two-step-away pairs:
  internet domains

- att.com
- ibm.com

log(degree) - log(rank) = -0.82
Solution# S.1

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

= friends of friends (F.O.F.)

internet domains

att.com

ibm.com

log(degree)

$-0.82$

$\approx 0.8$PB -> a data center(!)
Solution# S.1

- Q: So what?
- A1: # of two-step-away internet domains
  \( O(d_{\text{max}}^2) \approx 10M^2 \log(\text{rank}) \log(\text{degree}) \)

\( \sim 0.8\text{PB} \to \text{a data center(!)} \)

Such patterns -> New algorithms

Gaussian trap
Solution# S.2: Eigen Exponent $E$

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

$A \mathbf{x} = \lambda \mathbf{x}$

Exponent = slope

$E = -0.48$

Rank of decreasing eigenvalue

May 2001
Roadmap

• Introduction – Motivation

• Problem#1: Patterns in graphs
  – Static graphs
    • degree, diameter, eigen,
    • Triangles
  – 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?
  - 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 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_{max}^2)

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! - O(E)
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
• Part#1: Patterns in graphs
  – Static graphs
    • Power law degrees; eigenvalues; triangles
    • Anti-pattern: NO good cuts!
  – Time-evolving graphs
• ....
• Conclusions
Background: Graph cut problem

• Given a graph, and $k$
• Break it into $k$ (disjoint) communities
Graph cut problem

- Given a graph, and $k$
- Break it into $k$ (disjoint) communities
- (assume: block diagonal = ‘cavemen’ graph)
Many algo’s for graph partitioning

• METIS [Karypis, Kumar +]
• 2nd eigenvector of Laplacian
• Modularity-based [Girwan+Newman]
• Max flow [Flake+]
• ...
• ...
• ...
• ...
Strange behavior of min cuts

• Subtle details: next
  – Preliminaries: min-cut plots of ‘usual’ graphs


“Min-cut” plot

• Do min-cuts recursively.

\[
\text{Mincut size} = \sqrt{N}
\]

N nodes
“Min-cut” plot

- Do min-cuts recursively.

N nodes

New min-cut

log (mincut-size / #edges)

log (# edges)
“Min-cut” plot

- Do min-cuts recursively.

N nodes

New min-cut

Better cut

log (mincut-size / #edges)

Slope = -0.5

log (# edges)
"Min-cut" plot

log (mincut-size / #edges) vs log (# edges)

Slope $= -1/d$

For a d-dimensional grid, the slope is $-1/d$

For a random graph (and clique), the slope is 0
Experiments

• Datasets:
  – Google Web Graph: 916,428 nodes and 5,105,039 edges
  – Lucent Router Graph: Undirected graph of network routers from [www.isi.edu/scan/mercator/maps.html](http://www.isi.edu/scan/mercator/maps.html); 112,969 nodes and 181,639 edges
  – User ➔ Website Clickstream Graph: 222,704 nodes and 952,580 edges

“Min-cut” plot

• What does it look like for a real-world graph?

\[
\text{log (mincut-size / #edges)}
\]

\[
\text{log (# edges)}
\]
Experiments

- Used the METIS algorithm [Karypis, Kumar, 1995]

**Graph:**
- Google Web graph
- Values along the y-axis are averaged
- “lip” for large # edges
- Slope of -0.4, corresponds to a 2.5-dimensional grid!
Experiments

- Used the METIS algorithm [Karypis, Kumar, 1995]

- Google Web graph
- Values along the y-axis are averaged
- “lip” for large # edges
- Slope of -0.4, corresponds to a 2.5-dimensional grid!
Experiments

- Same results for other graphs too…

Lucent Router graph

Clickstream graph

Slope ~ -0.57

Slope ~ -0.45
Why no good cuts?

- Answer: self-similarity (few foils later)
Roadmap

• Introduction – Motivation
• Part#1: Patterns in graphs
  – Static graphs
  – Time-evolving graphs
  – Why so many power-laws?
• Part#2: Cascade analysis
• 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:**
  – \([\text{diameter} \sim O(N^{1/3})]\)
  – \(\text{diameter} \sim O(\log N)\)
  – \(\text{diameter} \sim O(\log \log N)\)

• What is happening in real data?
T.1 Evolution of the Diameter

• Prior work on Power Law graphs hints at slowly growing diameter:
  – [diameter $\sim O(N^{1/3})$]
  – 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
  \[ E(t+1) =? 2 \times E(t) \]

Say, \( k \) friends on average
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! \( \sim 3x \)
  - But obeying the "Densification Power Law"
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
  \[ E(t+1) = \text{?} \times E(t) \]
• A: over-doubled! \( \sim 3x \)
  – 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

\[
\begin{align*}
N(t) & \sim 1.66 \\
E(t) & \sim 0.0002 \times 1.66 \\
R^2 & = 0.99
\end{align*}
\]
## MORE Graph Patterns

<table>
<thead>
<tr>
<th>Static</th>
<th>Unweighted</th>
<th>Weighted</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>L01.</strong> Power-law degree distribution [Faloutsos et al. '99, Kleinberg et al. '99, Chakrabarti et al. '04, Newman '04]</td>
<td></td>
<td><strong>L10.</strong> Snapshot Power Law (SPL) [McGlohon et al. '08]</td>
</tr>
<tr>
<td><strong>L02.</strong> Triangle Power Law (TPL) [Tsourakakis '08]</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>L03.</strong> Eigenvalue Power Law (EPL) [Siganos et al. '03]</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>L04.</strong> Community structure [Flake et al. '02, Girvan and Newman '02]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dynamic</th>
<th>Unweighted</th>
<th>Weighted</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>L05.</strong> Densification Power Law (DPL) [Leskovec et al. '05]</td>
<td></td>
<td><strong>L11.</strong> Weight Power Law (WPL) [McGlohon et al. '08]</td>
</tr>
<tr>
<td><strong>L06.</strong> Small and shrinking diameter [Albert and Barabási '99, Leskovec et al. '05]</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>L07.</strong> Constant size 2nd and 3rd connected components [McGlohon et al. '08]</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>L08.</strong> Principal Eigenvalue Power Law ($\lambda_1$PL) [Akoglu et al. '08]</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>L09.</strong> Bursty/self-similar edge/weight additions [Gomez and Santonja '98, Gribble et al. '98, Crovella and</td>
<td></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>Static</th>
<th>Unweighted</th>
<th>Weighted</th>
</tr>
</thead>
<tbody>
<tr>
<td>✔️ 2. Triangle Power Law (TPL) [Tsourakakis '08]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>✔️ 3. Eigenvalue Power Law (EPL) [Siganos et al. '03]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>✔️ 4. Community structure [ Flake et al. '02, Girvan and Newman '02]</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dynamic</th>
<th>Unweighted</th>
<th>Weighted</th>
</tr>
</thead>
<tbody>
<tr>
<td>✔️ 5. Densification Power Law (DPL) [Leskovec et al. '05]</td>
<td></td>
<td>L11. Weight Power Law (WPL) [McGlohon et al. '08]</td>
</tr>
<tr>
<td>✔️ 6. Small and shrinking diameter [Albert and Barabási '99, Leskovec et al. '05]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>✔️ 7. Constant size 2nd and 3rd connected components [McGlohon et al. '08]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>✔️ 8. Principal Eigenvalue Power Law ($\lambda_1$PL) [Akoglu et al. '08]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>✔️ 9. Bursty/self-similar edge/weight additions [Gomez and Santonja '98, Gribble et al. '98, Crovella and</td>
<td></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>Static</th>
<th>Weighted</th>
</tr>
</thead>
<tbody>
<tr>
<td>3. Eigenvalue Power Law (EPL) (Vigano et al. ’03)</td>
<td></td>
</tr>
<tr>
<td>4. Community structure (Blake et al. ’02, Girvan and Newman ’02)</td>
<td></td>
</tr>
</tbody>
</table>

### Dynamic

<p>| |</p>
<table>
<thead>
<tr>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>105. Copy Power Law (CPL) (Leskovec et al. ’05)</td>
</tr>
<tr>
<td>106. Small and shrinking diameter (Albert and Barabasi ’99, Leskovec et al. ’05)</td>
</tr>
<tr>
<td>107. Constant size 2nd and 3rd connected components (McGlohon et al. ’08)</td>
</tr>
</tbody>
</table>


Roadmap

• Introduction – Motivation
• Part#1: Patterns in graphs
  – …
  – Why so many power-laws?
  – Why no ‘good cuts’?
• Part#2: Cascade analysis
• Conclusions
2 Questions, one answer

• Q1: why so many power laws
• Q2: why no ‘good cuts’?
2 Questions, one answer

• Q1: why so many power laws
• Q2: why no ‘good cuts’?
• A: Self-similarity = fractals = ‘RMAT’ ~ ‘Kronecker graphs’
20” intro to fractals

- Remove the middle triangle; repeat
- \( \rightarrow \) Sierpinski triangle
- (Bonus question - dimensionality?
  - \( >1 \) (inf. perimeter – \( (4/3)^\infty \))
  - \( <2 \) (zero area – \( (3/4)^\infty \))
20” intro to fractals

Self-similarity -> no char. scale
-> power laws, eg:
2x the radius,
3x the #neighbors \( nn(r) \)
\( nn(r) = C \ r^{\log_3/\log_2} \)
20” intro to fractals

Self-similarity -> no char. scale
-> power laws, eg:
2x the radius,
3x the #neighbors \( \text{nn}(r) \)
\[ \text{nn}(r) = C r^{\log 3/\log 2} \]
20” intro to fractals

Self-similarity -> no char. scale
-> power laws, eg:
2x the radius,
3x the #neighbors
\[ \text{nn} = C \cdot r^{\log_3 \log_2} \]

Reminder:
Densification P.L.
(2x nodes, ~3x edges)
20”” intro to fractals

Self-similarity -> no char. scale
-> power laws, eg:
2x the radius,
3x the #neighbors

\[ nn = C r^{\log_3/\log_2} \]

2x the radius,
4x neighbors

\[ nn = C r^{\log_4/\log_2} = C r^2 \]

Case'14  
(c) 2014, C. Faloutsos
20” intro to fractals

Self-similarity -> no char. scale
-> power laws, eg:
2x the radius,
3x the #neighbors

\[ nn = C r^{\log_3/\log_2} = 1.58 \]

Fractal dim.

2x the radius,
4x neighbors

\[ nn = C r^{\log_4/\log_2} = C r^2 \]
20” intro to fractals

Self-similarity -> no char. scale
-> power laws, eg:
2x the radius,
3x the #neighbors

\[ \text{nn} = C r^{\log_3/\log_2} \]

Fractal dim.

2x the radius,
4x neighbors

\[ \text{nn} = C r^{\log_4/\log_2} = C r^2 \]
How does self-similarity help in graphs?

• A: RMAT/Kronecker generators
  – With self-similarity, we get all power-laws, automatically,
  – And small/shrinking diameter
  – And ‘no good cuts’

R-MAT: A Recursive Model for Graph Mining,
by D. Chakrabarti, Y. Zhan and C. Faloutsos,
SDM 2004, Orlando, Florida, USA

Realistic, Mathematically Tractable Graph Generation
and Evolution, Using Kronecker Multiplication,
by J. Leskovec, D. Chakrabarti, J. Kleinberg,
and C. Faloutsos, in PKDD 2005, Porto, Portugal
Graph gen.: Problem dfn

• Given a growing graph with count of nodes $N_1$, $N_2$, ...
• Generate a realistic sequence of graphs that will obey all the patterns
  – Static Patterns
    S1 Power Law Degree Distribution
    S2 Power Law eigenvalue and eigenvector distribution
    Small Diameter
  – Dynamic Patterns
    T2 Growth Power Law (2x nodes; 3x edges)
    T1 Shrinking/Stabilizing Diameters
Kronecker Graphs

Adjacency matrix

\[ G_1 \]

\[ \begin{pmatrix}
  1 & 1 & 0 \\
  1 & 1 & 1 \\
  0 & 1 & 1 \\
\end{pmatrix} \]
Kronecker Graphs

Intermediate stage

Adjacency matrix

\[
\begin{pmatrix}
1 & 1 & 0 \\
1 & 1 & 1 \\
0 & 1 & 1
\end{pmatrix}
\]

\( G_1 \)
Kronecker Graphs

Intermediate stage

Adjacency matrix

\[
\begin{pmatrix}
1 & 1 & 0 \\
1 & 1 & 1 \\
0 & 1 & 1
\end{pmatrix}
\]

\(G_1\)

\[
\begin{pmatrix}
G_1 & G_1 & 0 \\
G_1 & G_1 & G_1 \\
0 & G_1 & G_1
\end{pmatrix}
\]

\(G_2 = G_1 \otimes G_1\)

Adjacency matrix
Kronecker Graphs

• Continuing multiplying with $G_1$ we obtain $G_4$ and so on …

$G_4$ adjacency matrix
Kronecker Graphs

- Continuing multiplying with \( G_1 \) we obtain \( G_4 \) and so on …

\[ G_4 \text{ adjacency matrix} \]
Kronecker Graphs

• Continuing multiplying with $G_1$ we obtain $G_4$ and so on …

$G_4$ adjacency matrix
Kronecker Graphs

- Continuing multiplying with $G_1$ we obtain $G_4$ and so on …

Holes within holes; Communities within communities

$G_4$ adjacency matrix
Properties:

- We can PROVE that
  - Degree distribution is multinomial ~ power law
  - Diameter: constant
  - Eigenvalue distribution: multinomial
  - First eigenvector: multinomial
Problem Definition

• Given a growing graph with nodes $N_1, N_2, ...$
• Generate a realistic sequence of graphs that will obey all the patterns
  – Static Patterns
    ✓ Power Law Degree Distribution
    ✓ Power Law eigenvalue and eigenvector distribution
    ✓ Small Diameter
  – Dynamic Patterns
    ✓ Growth Power Law
    ✓ Shrinking/Stabilizing Diameters
• First generator for which we can prove all these properties
Impact: Graph500

- Based on RMAT (= 2x2 Kronecker)
- Standard for graph benchmarks
- http://www.graph500.org/
- Competitions 2x year, with all major entities: LLNL, Argonne, ITC-U. Tokyo, Riken, ORNL, Sandia, PSC, …

To iterate is human, to recurse is divine
Roadmap

- Introduction – Motivation
- Part#1: Patterns in graphs
  - …
  - Q1: Why so many power-laws?
  - Q2: Why no ‘good cuts’?
- Part#2: Cascade analysis
- Conclusions

A: real graphs -> self similar -> power laws
Q2: Why ‘no good cuts’?

• A: self-similarity
  – Communities within communities within communities …
Kronecker Product – a Graph

- Continuing multiplying with $G_1$ we obtain $G_4$ and so on …

$G_4$ adjacency matrix
Kronecker Product – a Graph

- Continuing multiplying with $G_1$ we obtain $G_4$ and so on …

Communities within communities within communities …

$G_4$ adjacency matrix

‘Linux users’
‘Mac users’
‘win users’
Kronecker Product – a Graph

- Continuing multiplying with $G_1$ we obtain $G_4$ and so on …

How many Communities?
3?
9?
27?

$G_4$ adjacency matrix
Kronecker Product – a Graph

- Continuing multiplying with $G_1$ we obtain $G_4$ and so on …

Communities within communities within communities …

How many Communities?
3?
9?
27?

A: one – but not a typical, block-like community…

$G_4$ adjacency matrix
Communities?  (Gaussian) Clusters?  Piece-wise flat parts?

# songs

age

Case'14  (c) 2014, C. Faloutsos
Wrong questions to ask!

# songs

age

Wrong questions to ask!
Summary of Part#1

- *many* patterns in real graphs
  - Small & shrinking diameters
  - Power-laws everywhere
  - Gaussian trap
  - ‘no good cuts’

- Self-similarity (RMAT/Kronecker): good model
Part 2: Cascades & Immunization
Why do we care?

- Information Diffusion
- Viral Marketing
- Epidemiology and Public Health
- Cyber Security
- Human mobility
- Games and Virtual Worlds
- Ecology
- ........
Roadmap

• Introduction – Motivation
• Part#1: Patterns in graphs
• Part#2: Cascade analysis
  – (Fractional) Immunization
  – Epidemic thresholds
• Conclusions
Fractional Immunization of Networks
B. Aditya Prakash, Lada Adamic, Theodore Iwashyna (M.D.), Hanghang Tong, Christos Faloutsos

SDM 2013, Austin, TX
Whom to immunize?

- Dynamical Processes over networks

- Each circle is a hospital
- ~3,000 hospitals
- More than 30,000 patients transferred

Problem: Given $k$ units of disinfectant, whom to immunize?
Whom to immunize?

~6x fewer!

CURRENT PRACTICE

OUR METHOD

Hospital-acquired inf.: 99K+ lives, $5B+ per year
Fractional Asymmetric Immunization

Drug-resistant Bacteria (like XDR-TB)

Hospital

Another Hospital

(c) 2014, C. Faloutsos
Fractional Asymmetric Immunization

Hospital

Another Hospital

Case'14
(c) 2014, C. Faloutsos
Fractional Asymmetric Immunization

Hospital

Another Hospital

(c) 2014, C. Faloutsos
Fractional Asymmetric Immunization

Problem:
Given $k$ units of disinfectant, distribute them to maximize hospitals saved
Fractional Asymmetric Immunization

**Problem:**
Given \( k \) units of disinfectant, distribute them to maximize hospitals saved @ 365 days
Straightforward solution:

Simulation:
1. Distribute resources
2. ‘infect’ a few nodes
3. Simulate evolution of spreading
   - (10x, take avg)
4. Tweak, and repeat step 1
Straightforward solution:

Simulation:
1. Distribute resources
2. ‘infect’ a few nodes
3. Simulate evolution of spreading
   – (10x, take avg)
4. Tweak, and repeat step 1
Straightforward solution:

Simulation:
1. Distribute resources
2. ‘infect’ a few nodes
3. Simulate evolution of spreading
   – (10x, take avg)
4. Tweak, and repeat step 1
Straightforward solution:

Simulation:

1. Distribute resources
2. ‘infect’ a few nodes
3. Simulate evolution of spreading
   - (10x, take avg)
4. Tweak, and repeat step 1
Running Time

- Simulations: > 1 week
- SMART-ALLOC: 14 secs

> 30,000x speed-up!

Wall-Clock Time:

better
Experiments

# infected

\[ K = 120 \]

# epochs

SMART-ALLOC

uniform

better
What is the ‘silver bullet’?

A: Try to decrease connectivity of graph

Q: how to measure connectivity?
   – Avg degree? Max degree?
   – Std degree / avg degree?
   – Diameter?
   – Modularity?
   – ‘Conductance’ (~min cut size)?
   – Some combination of above?
What is the ‘silver bullet’?

A: Try to decrease connectivity of graph

Q: how to measure connectivity?
A: first eigenvalue of adjacency matrix

Q1: why??
(Q2: dfn & intuition of eigenvalue ? )
Why eigenvalue?

A1: ‘G2’ theorem and ‘eigen-drop’:

• For (almost) any type of virus
• For any network
• -> no epidemic, if small-enough first eigenvalue ($\lambda_1$) of adjacency matrix

Threshold Conditions for Arbitrary Cascade Models on Arbitrary Networks, B. Aditya Prakash, Deepayan Chakrabarti, Michalis Faloutsos, Nicholas Valler, Christos Faloutsos, ICDM 2011, Vancouver, Canada
Why eigenvalue?

A1: ‘G2’ theorem and ‘eigen-drop’:

• For (almost) **any** type of virus
• For **any** network
• -> no epidemic, if small-enough first eigenvalue ($\lambda_1$) of adjacency matrix

• Heuristic: for immunization, try to min $\lambda_1$
• The smaller $\lambda_1$, the closer to extinction.
G2 theorem

Threshold Conditions for Arbitrary Cascade Models on Arbitrary Networks
B. Aditya Prakash, Deepayan Chakrabarti, Michalis Faloutsos, Nicholas Valler, Christos Faloutsos
IEEE ICDM 2011, Vancouver

extended version, in arxiv
http://arxiv.org/abs/1004.0060
~10 pages proof
Our thresholds for some models

- \( s = \text{effective strength} \)
- \( s < 1 : \text{below threshold} \)

<table>
<thead>
<tr>
<th>Models</th>
<th>Effective Strength (s)</th>
<th>Threshold (tipping point)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIS, SIR, SIRS, SEIR</td>
<td>( s = \lambda ) \left( \frac{\beta}{\delta} \right)</td>
<td></td>
</tr>
<tr>
<td>SIV, SEIV</td>
<td>( s = \lambda ) \left( \frac{\beta \gamma}{\delta (\gamma + \theta)} \right)</td>
<td>( s = 1 )</td>
</tr>
<tr>
<td>SI_1I_2V_1V_2 (H.I.V.)</td>
<td>( s = \lambda ) \left( \frac{\beta_1v_2 + \beta_2 \varepsilon}{v_2 (\varepsilon + \nu_1)} \right)</td>
<td></td>
</tr>
</tbody>
</table>
Our thresholds for some models

- \( s = \text{effective strength} \)
- \( s < 1 : \text{below threshold} \)

<table>
<thead>
<tr>
<th>Model</th>
<th>Effective Strength</th>
<th>Threshold (tipping point)</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIS, SIR, SIRS, SEIR</td>
<td>( s = \lambda ) ( \begin{pmatrix} \beta \ \delta \end{pmatrix} )</td>
<td></td>
</tr>
<tr>
<td>SIV, SEIV</td>
<td>( s = \lambda . \left( \frac{\beta \gamma}{\delta (\gamma + \theta)} \right) )</td>
<td>( s = 1 )</td>
</tr>
<tr>
<td>SI_1I_2V_1V_2 (H.I.V.)</td>
<td>( s = \lambda . \left( \frac{\beta_1 v_2 + \beta_2 \varepsilon}{v_2 (\varepsilon + v_1)} \right) )</td>
<td></td>
</tr>
</tbody>
</table>
Roadmap

• Introduction – Motivation
• Part#1: Patterns in graphs
• Part#2: Cascade analysis
  – (Fractional) Immunization
  – intuition behind $\lambda_1$
• Conclusions
Intuition for $\lambda$

“Official” definitions:

- Let $A$ be the adjacency matrix. Then $\lambda$ is the root with the largest magnitude of the characteristic polynomial of $A$ [$\det(A - xI)$].
- Also: $Ax = \lambda x$

Neither gives much intuition!

“Un-official” Intuition

- For ‘homogeneous’ graphs, $\lambda == \text{degree}$
- $\lambda \sim \text{avg degree}$
  - done right, for skewed degree distributions
Largest Eigenvalue ($\lambda$)

$$\lambda \approx 2$$

(a) Chain

$$\lambda = \sqrt{N}$$

(b) Star

$$\lambda = N-1$$

(c) Clique

$N = 1000$ nodes

(c) 2014, C. Faloutsos
Largest Eigenvalue ($\lambda$)

better connectivity $\rightarrow$ higher $\lambda$

\( \lambda \approx 2 \)
(a) Chain

\( \lambda = \sqrt{N} \)
(b) Star

\( \lambda = N-1 \)
(c) Clique

\( N = 1000 \) nodes

\( \lambda \approx 2 \)

\( \lambda = 31.67 \)

\( \lambda = 999 \)

Case'14

(c) 2014, C. Faloutsos
Examples: Simulations – SIR (mumps)

(a) Infection profile
(b) “Take-off” plot

PORTLAND graph: synthetic population,
31 million links, 6 million nodes
Examples: Simulations – SIRS (pertussis)

(a) Infection profile

PORTLAND graph: synthetic population, 31 million links, 6 million nodes

(b) “Take-off” plot
Immunization - conclusion

In (almost any) immunization setting,

• Allocate resources, such that to

• Minimize $\lambda_1$

• (regardless of virus specifics)

• Conversely, in a market penetration setting
  – Allocate resources to
  – Maximize $\lambda_1$
Roadmap

• Introduction – Motivation
• Part#1: Patterns in graphs
• Part#2: Cascade analysis
  – (Fractional) Immunization
  – Epidemic thresholds

• What next?
• Acks & Conclusions
• [Tools: ebay fraud; tensors; spikes]
Challenge #1: ‘Connectome’ – brain wiring

- Which neurons get activated by ‘bee’
- How wiring evolves
- Modeling epilepsy
Challenge#2: Time evolving networks / tensors

• Periodicities? Burstiness?
• What is ‘typical’ behavior of a node, over time
• Heterogeneous graphs (= nodes w/ attributes)
Roadmap

• Introduction – Motivation
• Part#1: Patterns in graphs
• Part#2: Cascade analysis
  – (Fractional) Immunization
  – Epidemic thresholds
• Acks & Conclusions
• [Tools: ebay fraud; tensors; spikes]

Off line
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
Project info: PEGASUS

www.cs.cmu.edu/~pegasus

Results on large graphs: with Pegasus + hadoop + M45

Apache license

Code, papers, manual, video

Prof. U Kang

Prof. Polo Chau
Cast

Akoglu, Leman

Beutel, Alex

Chau, Polo

Kang, U

Koutra, Danai

Lee, Jay Yoon

Prakash, Aditya

Papalexakis, Vagelis

Shah, Neil

Tong, Hanghang
CONCLUSION#1 – Big data

- **Large** datasets reveal patterns/outliers that are invisible otherwise
CONCLUSION#2 – self-similarity

- powerful tool / viewpoint
  - Power laws; shrinking diameters
  - Gaussian trap (eg., F.O.F.)
  - ‘no good cuts’
  - RMAT – graph500 generator
CONCLUSION#3 – eigen-drop

- Cascades & immunization: G2 theorem & eigenvalue

CURRENT PRACTICE

OUR METHOD

~6x fewer!

[US-MEDICARE NETWORK 2005]

> 30,000x speed-up!

14 secs
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
QUESTIONS?

Cross-disciplinarity

CMU SCS
TAKE HOME MESSAGE:

Case'14  (c) 2014, C. Faloutsos  128