15-826: Multimedia Databases and Data Mining

Project lecture #1: Graph mining - patterns
Christos Faloutsos

Must-read Material – 1-of-2

  - Part I (patterns)
Must-read Material 2-of-2

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

Problem

• Are real graphs random?
Conclusions

• Are real graphs random?
• NO!
  – Static patterns
    • Small diameters
    • Skewed degree distribution
    • Shrinking diameters
  – Weighted
  – Time-evolving

• Many power laws – log-logistic
• Take logarithms
• Time-evolving
Main outline

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

Outline

- Introduction – Motivation
- Problem: Patterns in graphs
- Problem#2: Scalability
- Conclusions
Graphs - why should we care?

• IR: bi-partite graphs (doc-terms)
  
  \[ \begin{align*} 
  D_1 & \quad \cdots \quad D_N \\
  T_1 & \quad \cdots \quad T_M 
  \end{align*} \]

• 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: 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?

---

- anomalies (rarities) <-> 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?
  – anomalies (rarities) \leftrightarrow \textbf{patterns}
  – \textbf{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 (‘6 degrees’, ‘Kevin Bacon’)
  – 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(rank) vs. log(degree)

att.com

ibm.com
Solution# S.1

• Power law in the degree distribution
  [SIGCOMM99]

  internet domains

  ![Graph showing log(rank) vs log(degree)]
  
  -0.82

- Q: So what?

  internet domains

  ![Graph showing log(rank) vs log(degree)]
  
  -0.82
Solution# S.1

• Q: So what?
• A1: # of two-step-away pairs: internet domains

Q: So what?
A1: # of two-step-away pairs: 100^2 * N = 10 Trillion internet domains

\[ \text{att.com} \]
\[ \text{ibm.com} \]

log(rank)

\( \log(\text{rank}) = 0.82 \)

\( \log(\text{degree}) \)
Solution# S.1

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

\[ \log(\text{rank}) = \log(\text{degree}) - 0.82 \]

- IBM.com
- ATT.com

15-826
(c) C. Faloutsos, 2019

Solution# S.1

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

\[ \log(\text{rank}) = \log(\text{degree}) - 0.82 \]

- IBM.com
- ATT.com

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

DCO @ CMU

15-826
(c) C. Faloutsos, 2019
Solution# S.1

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

Such patterns -> New algorithms

Observation – big-data:

- \(O(N^2)\) algorithms are \(~\text{intractable}~\) - \(N=1B\)
- \(N^2\) seconds = 31B years (>2x age of universe)
Observation – big-data:

- O($N^2$) algorithms are ~intractable - N=1B
- $N^2$ seconds = 31B years
- 1,000 machines

(c) C. Faloutsos, 2019

Observation – big-data:

- O($N^2$) algorithms are ~intractable - N=1B
- $N^2$ seconds = 31B years
- 1M machines

(c) C. Faloutsos, 2019
Observation – big-data:

• $O(N^2)$ algorithms are ~intractable - $N=1B$

And parallelism might not help

• $N^2$ seconds = 31B years
• 10B machines ~ $10$Trillion

(c) C. Faloutsos, 2019
Solution# S.2: Eigen Exponent $E$

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

\[ A \mathbf{x} = \lambda \mathbf{x} \]

- Eigenvalue
- Exponent = slope
- $E = -0.48$
- May 2001

[Mihail, Papadimitriou ’02]: slope is $\frac{1}{2}$ of rank exponent

15-826  (c) C. Faloutsos, 2019
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]

(count) degree

trusts-2000-people user

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’
List of Static Patterns

- S.1 degree
- S.2 eigenvalues
- S.3 small diameter
- S.4/5 Triangle laws
- (S.6) NLCC non-largest conn. components
- (S.7) eigen plots
- (S.8) radius plot

In textbook

S.3 small diameters

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

- S.1 degree
- S.2 eigenvalues
- S.3 small diameter
  - S.4/5 Triangle laws
  - (S.6) NLCC non-largest conn. components
  - (S.7) eigen plots
  - (S.8) radius plot

Solution# S.4: Triangle ‘Laws’

- Real social networks have a lot of triangles
Solution# S.4: Triangle ‘Laws’

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

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

HEP-TH

ASN

Epinions

X-axis: # of participating triangles
Y: count (~ pdf)
Triangle Law: #S.4
[Tsourakakis ICDM 2008]

HEP-TH

ASN

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

15-826
10^2

Triangle Law: #S.5
[Tsourakakis ICDM 2008]

Reuters

SN

Epinions

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

15-826
outsos, 2019
Triangle Law: Computations  
[Tsourakakis ICDM 2008]

But: triangles are expensive to compute  
(3-way join; several approx. algos)  
Q: Can we do that quickly?

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

[Tsourakakis ICDM 2008]
Wikipedia graph 2006-Nov-04
≈ 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]
Triangle counting for large graphs?

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

(c) C. Faloutsos, 2019
Triangle counting for large graphs?

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

(c) C. Faloutsos, 2019
List of Static Patterns

- S.1 degree
- S.2 eigenvalues
- S.3 small diameter
- S.4/5 Triangle laws
- (S.6) NLCC non-largest conn. components
- (S.7) eigen plots
- (S.8) radius plot

In textbook

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).
S.6: NLCC

- Connected Components – 4 observations:

1) 10K x larger than next
S.6: NLCC

- Connected Components

Count

2) \(~0.7\)B singleton nodes

3) SLOPE!
S.6: NLCC

• Connected Components

4) Spikes!

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

Why?

suspicious financial-advice sites
(not existing now)

(c) C. Faloutsos, 2019
S.6: persists over time

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

Stable tail slope after the gelling point

List of Static Patterns

- S.1 degree
- S.2 eigenvalues
- S.3 small diameter
- S.4/5 Triangle laws
- (S.6) NLCC non-largest conn. components
- (S.7) eigen plots
- (S.8) radius plot

In textbook
EigenSpokes


**Useful for fraud detection!**

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

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

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

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

(c) C. Faloutsos, 2019
EigenSpokes

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

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

(c) C. Faloutsos, 2019
EigenSpokes

- EE plot:
- Scatter plot of scores of $u_1$ vs $u_2$
- One would expect
  - Many points @ origin
  - A few scattered ~randomly

(c) C. Faloutsos, 2019
EigenSpokes - pervasiveness

- Present in mobile social graph
  - across time and space

- Patent citation graph

EigenSpokes - explanation

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

Near-cliques, or near-bipartite-cores, loosely connected
**EigenSpokes - explanation**

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!

magnified bipartite community

Useful for fraud detection!
Bipartite Communities!

IP – port scanners

victims

Useful for fraud detection!

List of Static Patterns

- S.1 degree
- S.2 eigenvalues
- S.3 small diameter
- S.4/5 Triangle laws
- (S.6) NLCC non-largest conn. components
- (S.7) eigen plots
- (S.8) radius plot

In textbook
**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** (\(N\sim 1B\))

- Our HADI: linear on \(E (~10B)\)
  - Near-linear scalability wrt # machines
  - Several optimizations -> 5x faster

---

**Count**

- 19+ [Barabasi+]
- \(~1999, \sim 1M\) nodes
YahooWeb graph (120Gb, 1.4B nodes, 6.6 B edges)
- Largest publicly available graph ever studied.

19+ [Barabasi+]
- ~1999, ~1M nodes

14 (dir.)
- ~7 (undir.)

19+? [Barabasi+]

(c) C. Faloutsos, 2019
YahooWeb graph (120Gb, 1.4B nodes, 6.6B edges)
• 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.

List of Static Patterns
- S.1 degree
- S.2 eigenvalues
- S.3 small diameter
- S.4/5 Triangle laws
- (S.6) NLCC non-largest conn. components
- (S.7) eigen plots
- (S.8) radius plot
- Other observations / patterns?
List of Static Patterns

- S.1 degree
- S.2 eigenvalues
- S.3 small diameter
- S.4/5 Triangle laws
- (S.6) NLCC non-largest conn. components
- (S.7) eigen plots
- (S.8) radius plot
- Other observations / patterns?

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’?

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

Outline

- Introduction – Motivation
- Problem: 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?*
Observation W.1: Fortification

More donors, more $ ?

‘Reagan’

\$10

\$5

‘Clinton’

\$7

15-826

(c) C. Faloutsos, 2019

Observation W.1: fortification: Snapshot Power Law

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

More donors, even more $

\$10

\$5

In-weights ($)

15-826

(c) C. Faloutsos, 2019

Orgs-Candidates

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

Edges (# donors)
Outline

• Introduction – Motivation
• Problem: 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)
List of Dynamic Patterns

• D.1 diameter
• D.2 densification
• D.3 gelling point
• D.4 NLCC over time
• D.5 Eigenvalue over time
• D.6 Popularity over time
• D.7 phonecall duration

D.1 Evolution of the Diameter

• Prior work on Power Law graphs hints at \textit{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?
D.1 Evolution of the Diameter

- Prior work on Power Law graphs hints at slowly growing diameter:
  - [diameter ~ O(N^{1/3})]
  - diameter ~ O(log N)
  - diameter ~ O(log log N)
- What is happening in real data?
- Diameter shrinks over time

D.1 Diameter – “Patents”

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

- D.1 diameter
- D.2 densification
- D.3 gelling point
- D.4 NLCC over time
- D.5 Eigenvalue over time
- D.6 Popularity over time
- D.7 phonecall duration

In textbook

D.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
  \[ E(t+1) =? 2 \times E(t) \]
D.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 $E(t+1) =$? $2 \times E(t)$
- A: over-doubled!

- But obeying the "Densification Power Law"

D.2 Densification – Patent Citations

- Citations among patents granted
- @1999
  - 2.9 M nodes
  - 16.5 M edges
- Each year is a datapoint
List of Dynamic Patterns

- D.1 diameter
- D.2 densification
  - D.3 gelling point
  - D.4 NLCC over time
  - D.5 Eigenvalue over time
  - D.6 Popularity over time
  - D.7 phonecall duration

In textbook

More on Time-evolving graphs

M. McGlohon, L. Akoglu, and C. Faloutsos

*Weighted Graphs and Disconnected Components: Patterns and a Generator.*

*SIG-KDD 2008*
D.3 Gelling Point

- Diameter, over time

Most real graphs display a gelling point
- After gelling point, they exhibit typical behavior. This is marked by a spike in diameter.
**D.3 Gelling Point**

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

![Graph showing IMDB diameter over time at t=1914]

**List of Dynamic Patterns**

- D.1 diameter
- D.2 densification
- D.3 gelling point
- D.4 NLCC over time
- D.5 Eigenvalue over time
- D.6 Popularity over time
- D.7 phonecall duration

*In textbook*
Observation D.4: 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 D.4: NLCC behavior

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

(``NLCC’’ = non-largest conn. components)

YES – Do they continue to grow in size?
YES – or do they shrink?
YES – or stabilize?

• After the gelling point, the GCC takes off, but NLCC’s remain ~constant (actually, oscillate).
List of Dynamic Patterns

- D.1 diameter
- D.2 densification
- D.3 gelling point
- D.4 NLCC over time
  - D.5 Eigenvalue over time
  - D.6 Popularity over time
  - D.7 phonecall duration

Timing for Blogs

Cascading Behavior in Large Blog Graphs: Patterns and a model
Jure Leskovec, Mary McGlohon, Christos Faloutsos, Natalie Glance, Matthew Hurst
SDM’07
D.6 : popularity over time

Post popularity drops-off – exponentially?

POWER LAW!
Exponent?
D.6 : 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

-1.5 slope

J. G. Oliveira & A.-L. Barabási Human Dynamics: The Correspondence Patterns of Darwin and Einstein. 
*Nature* 437, 1251 (2005) . [PDF]
List of Dynamic Patterns

- D.1 diameter
- D.2 densification
- D.3 gelling point
- D.4 NLCC over time
  - D.5 Eigenvalue over time
- D.6 Popularity over time
- D.7 phonecall duration

D.7: duration of phonecalls

*Surprising Patterns for the 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 (?)

No Power Law!
‘TLaC: Lazy Contractor’

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

Odds ratio =
\[
\frac{\text{Casualties(<x)}}{\text{Survivors(>=x)}}
\]

== power law

Log-logistic distribution

• CDF(t)/(1 - CDF(t)) == OR(t)
• For log-logistic: \(\log[\text{OR}(t)] = \beta + \rho \log(t)\)

Odds ratio =
\[
\frac{\text{Casualties(<x)}}{\text{Survivors(>=x)}}
\]

== power law
Log-logistic distribution

- CDF(t)/(1 - CDF(t)) = OR(t)
- For log-logistic: log[OR(t)] = β + ρ*log(t)

- PDF looks like hyperbola;
- and, if clipped, like power-law
Log-logistic distribution

- Logistic distribution: CDF -> sigmoid
- LOG-Logistic distribution: $x \rightarrow \ln(x)$

CDF$(x) = \frac{1}{1+\exp(-x)}$  
CDF$(x) = \frac{1}{1+1/x}$
Log-logistic distribution

Nice 1 page description: section II of

Pravallika Devineni, Danai Koutra, Michalis Faloutsos, and Christos Faloutsos.
*If walls could talk: Patterns and anomalies in Facebook wallposts.*
*ASONAM 2015, pp 367-374.*
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)
Outliers:

Conclusions

- Are real graphs random?
- NO!
  - Static patterns
    - Small diameters
    - Skewed degree distribution
    - Shrinking diameters
  - Weighted
  - Time-evolving
Conclusions

• Are real graphs random?
• NO!
  – Static patterns
    • Small diameters
    • Skewed degree distribution
  – Weighted
  – Time-evolving

• Many power laws – log-logistic
• Take logarithms

Next lecture:

• Anomaly detection tools (OddBall, etc)