Graph analysis: laws & tools

Christos Faloutsos
Carnegie Mellon University

Overall Outline

- Laws (mainly, power laws)
- Generators and
- Tools

Outline

- Problem definition / Motivation
- Static & dynamic laws; generators
- Tools: CenterPiece graphs; Tensors
- Other projects (Virus propagation, e-bay fraud detection)
- Conclusions
Motivation

Data mining: ~ find patterns (rules, outliers)
• Problem#1: How do real graphs look like?
• Problem#2: How do they evolve?
• Problem#3: How to generate realistic graphs

TOOLS
• Problem#4: Who is the ‘master-mind’?
• Problem#5: Track communities over time

Problem#1: Joint work with
Dr. Deepayan Chakrabarti
(CMU/Yahoo R.L.)

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

Problem #1 - network and graph mining

- How does the Internet look like?
- How does the web look like?
- What is ‘normal’/‘abnormal’?
- which patterns/laws hold?
Graph mining

• Are real graphs random?

Laws and patterns

• Are real graphs random?
• A: NO!!
  – Diameter
  – in- and out- degree distributions
  – other (surprising) patterns

Solution#1

• Power law in the degree distribution
  [SIGCOMM99]

[Diagram showing log(degree) vs. log(rank) with points at att.com and ibm.com]
Solution#1': Eigen Exponent $E$

Eigenvalue

Exponent = slope
$E = -0.48$

May 2001

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

But:

How about graphs from other domains?

Web

- In- and out-degree distribution of web sites
  [Barabasi], [IBM-CLEVER]

log indegree

from [Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, Andrew Tomkins]
Web

- In- and out-degree distribution of web sites [Barabasi, IBM-CLEVER]

log(freq)

from [Ravi Kumar, Prabhakar Raghavan, Sridhar Rajagopalan, Andrew Tomkins]

log indegree

The Peer-to-Peer Topology

- Frequency versus degree
- Number of adjacent peers follows a power-law

More power laws:

citation counts: (citeseer.nj.nec.com 6/2001)
Swedish sex-web

Nodes: people (Females; Males)

Links: sexual relationships

4781 Swedes; 18-74;
59% response rate.

Liljeros et al. Nature 2001

Swedish sex-web

Nodes: people (Females; Males)

Links: sexual relationships

4781 Swedes; 18-74;
59% response rate.

Liljeros et al. Nature 2001

More power laws:

• web hit counts [w/ A. Montgomery]
**Motivation**

Data mining: ~ find patterns (rules, outliers)

✓ Problem#1: How do real graphs look like?
• Problem#2: How do they evolve?
• Problem#3: How to generate realistic graphs

TOOLS
• Problem#4: Who is the ‘master-mind’?
• Problem#5: Track communities over time

**Outline**

• Problem definition / Motivation
  ➔ Static & dynamic laws; generators
  • Tools: CenterPiece graphs; Tensors
  • Other projects (Virus propagation, e-bay fraud detection)
  • Conclusions

**epinions.com**

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

(out) degree

- trusts-2000-people user
Problem#2: Time evolution

• with Jure Leskovec (CMU/MLD)

• and Jon Kleinberg (Cornell – sabb. @ CMU)

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?

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
Diameter – ArXiv citation graph

- Citations among physics papers
- One graph per year
- 2003:
  - 29,555 papers,
  - 352,807 citations

Diameter – “Autonomous Systems”

- Graph of Internet
- One graph per day
- 1997 – 2000
- 2000
  - 6,000 nodes
  - 26,000 edges

Diameter – “Affiliation Network”

- Graph of collaborations in physics – authors linked to papers
- 10 years of data
- 2002
  - 60,000 nodes
    - 20,000 authors
    - 38,000 papers
  - 133,000 edges
Diameter – “Patents”

- Patent citation network
- 25 years of data
- 1999
  - 2.9 million nodes
  - 16.5 million edges

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
  - $E(t+1) = ? 2 \times E(t)$

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”
Densification – Physics Citations

- Citations among physics papers
- 2003:
  - 29,555 papers, 352,807 citations

\[ N(t) \]
\[ E(t) \]

\[ 1.69 \]

1: tree
**Densification – Physics Citations**

- Citations among physics papers
- 2003:
  - 29,555 papers, 352,807 citations

**Densification – Patent Citations**

- Citations among patents granted
- 1999
  - 2.9 million nodes
  - 16.5 million edges
- Each year is a datapoint

**Densification – Autonomous Systems**

- Graph of Internet
- 2000
  - 6,000 nodes
  - 26,000 edges
- One graph per day
Densification – Affiliation Network

- Authors linked to their publications
- 2002
  - 60,000 nodes
    - 20,000 authors
    - 38,000 papers
  - 133,000 edges

Motivation

Data mining: ~ find patterns (rules, outliers)

- Problem#1: How do real graphs look like?
- Problem#2: How do they evolve?
- Problem#3: How to generate realistic graphs

TOOLS

- Problem#4: Who is the ‘master-mind’?
- Problem#5: Track communities over time

Outline

- Problem definition / Motivation
  - Static & dynamic laws; generators
    - Tools: CenterPiece graphs; Tensors
    - Other projects (Virus propagation, e-bay fraud detection)
  - Conclusions
Problem Definition

• Given a growing graph with count of nodes \( N_1, N_2, \ldots \)
• 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

Idea: Self-similarity

• Leads to power laws
• Communities within communities
• …
Kronecker Product – a Graph

- Continuing multiplying with $G_j$ we obtain $G_d$ and so on …
Kronecker Product – a Graph

- Continuing multiplying with $G_j$, we obtain $G_d$ and so on …

$$G_g \text{ adjacency matrix}$$

Properties:

- We can prove that
  - Degree distribution is multinomial ~ power law
  - Diameter: constant
  - Eigenvalue distribution: multinomial
  - First eigenvector: multinomial
- See [Leskovec+, PKDD’05] for proofs

Problem Definition

- Given a growing graph with nodes $N_x$, $N_x$ …
- Generate a realistic sequence of graphs that will obey all the patterns
  - Static Patterns
    - Power Law Degree Distribution
    - Power Law eigenvector and eigenvector distribution
    - Small Diameter
  - Dynamic Patterns
    - Growth Power Law
    - Shrinking/Stabilizing Diameters
- First and only generator for which we can prove all these properties
Stochastic Kronecker Graphs

- Create \( N \times N \) probability matrix \( P \).
- Compute the \( k^{th} \) Kronecker power \( P^k \).
- For each entry \( p_{uv} \) of \( P^k \) include an edge \((u,v)\) with probability \( p_{uv} \).

### Example

- \( P_1 \):
  - \[ \begin{array}{cccc} 0.4 & 0.2 & 0.1 & 0.3 \\ 0.16 & 0.08 & 0.08 & 0.04 \\ 0.04 & 0.12 & 0.02 & 0.06 \\ 0.04 & 0.02 & 0.12 & 0.06 \\ 0.01 & 0.03 & 0.03 & 0.09 \end{array} \]  
- \( P_k \):
  - \( k \)th Kronecker power

### Instance

- Matrix \( G \):
  - \[ \begin{array}{cccc} 0.09 & 0.03 & 0.03 & 0.01 \\ 0.06 & 0.12 & 0.02 & 0.04 \\ 0.06 & 0.02 & 0.12 & 0.04 \\ 0.04 & 0.08 & 0.08 & 0.16 \end{array} \]

### Experiments

- How well can we match real graphs?
  - **Arxiv**: physics citations:
    - 30,000 papers, 350,000 citations
    - 10 years of data
  - **U.S. Patent citation network**
    - 4 million patents, 16 million citations
    - 37 years of data
  - **Autonomous systems – graph of internet**
    - Single snapshot from January 2002
    - 6,400 nodes, 26,000 edges
- We show both static and temporal patterns

### Arxiv – Degree Distribution

- **Real graph**
- **Deterministic Kronecker**
- **Stochastic Kronecker**
(Q: how to fit the parm’s?)

A:
- Stochastic version of Kronecker graphs +
- Max likelihood +
- Metropolis sampling
- [Leskovec+, ICML’07]

Experiments on real AS graph

Degree distribution

Hop plot

Adjacency matrix eigen values

Network value

Conclusions

- Kronecker graphs have:
  - All the static properties
    - Heavy tailed degree distributions
    - Small diameter
    - Multinomial eigenvalues and eigenvectors
  - All the temporal properties
    - Densification Power Law
    - Shrinking/Stabilizing Diameters
  - We can formally prove these results
Outline

- Problem definition / Motivation
- Static & dynamic laws; generators
  - Tools: CenterPiece graphs; Tensors
- Other projects (Virus propagation, e-bay fraud detection)
- Conclusions

Motivation

Data mining: ~ find patterns (rules, outliers)

✓ Problem#1: How do real graphs look like?
✓ Problem#2: How do they evolve?
✓ Problem#3: How to generate realistic graphs

TOOLS

✓ Problem#4: Who is the ‘master-mind’?
  - Problem#5: Track communities over time

Problem#4: MasterMind – ‘CePS’

- w/ Hanghang Tong, KDD 2006
- htong <at> cs.cmu.edu
Center-Piece Subgraph (CePS)

- **Given** Q query nodes
- **Find** Center-piece \( \leq b \)

- App.
  - Social Networks
  - Law Enforcement, …

- Idea:
  - Proximity \( \rightarrow \) random walk with restarts

---

Center-Piece Subgraph (Ceps)

- **Given** Q query nodes
- **Find** Center-piece \( \leq b \)

- App.
  - Social Networks
  - Law Enforcement, …

- Idea:
  - Proximity \( \rightarrow \) random walk with restarts

---

Case Study: AND query

R. Agrawal
Jiawei Han
V. Vapnik
M. Jordan
Conclusions

- Q1: How to measure the importance?
  - A1: RWR+K_SoftAnd
- Q2: How to find connection subgraph?
  - A2: "Extract" Alg.
- Q3: How to do it efficiently?
  - A3: Graph Partition (Fast CePS)
    - ~90% quality
    - 6:1 speedup; 150x speedup (ICDM’06, b.p. award)

Outline

- Problem definition / Motivation
- Static & dynamic laws; generators
  - Tools: CenterPiece graphs; Tensors
  - Other projects (Virus propagation, e-bay fraud detection)
- Conclusions

Motivation

Data mining: ~ find patterns (rules, outliers)
- Problem#1: How do real graphs look like?
- Problem#2: How do they evolve?
- Problem#3: How to generate realistic graphs

TOOLS
- Problem#4: Who is the ‘master-mind’?
- Problem#5: Track communities over time
Tensors for time evolving graphs

- [Jimeng Sun+ KDD’06]
- [“”, SDM’07]
- [CF, Kolda, Sun, SDM’07 and SIGMOD’07 tutorial]

Social network analysis

- Static: find community structures

Social network analysis

- Static: find community structures
- Dynamic: monitor community structure evolution; spot abnormal individuals; abnormal time-stamps
Application 1: Multiway latent semantic indexing (LSI)

- Projection matrices specify the clusters
- Core tensors give cluster activation level

Crash course

- On SVD / spectral methods
- And tensors

SVD as spectral decomposition

\[ A \approx U \Sigma V^T = \sum_{i=1}^{k} \sigma_i u_i \circ v_i \]

- Best rank-k approximation in L2 and Frobenius
- SVD only works for static matrices (a single 2nd order tensor)

See also PARAFAC
### SVD - Example

- **A = U Σ V^T** - example:

  \[
  \begin{bmatrix}
  1 & 1 & 1 & 0 & 0 \\
  2 & 2 & 2 & 0 & 0 \\
  1 & 1 & 1 & 0 & 0 \\
  5 & 5 & 5 & 0 & 0 \\
  0 & 0 & 0 & 2 & 2 \\
  0 & 0 & 0 & 3 & 3 \\
  0 & 0 & 0 & 0 & 1 \\
  0 & 0 & 0 & 0 & 1 \\
  \end{bmatrix}
  \begin{bmatrix}
  0.18 & 0.36 & 0.18 & 0.90 & 0.53 \\
  0.80 & 0.27 & 0.58 & 0.58 & 0.71 \\
  \end{bmatrix}
  \begin{bmatrix}
  9.64 & 0 \\
  0 & 5.29 \\
  \end{bmatrix}
  \begin{bmatrix}
  0.58 & 0.58 & 0.71 \\
  0 & 0 & 0.71 \\
  \end{bmatrix}
  \]

- **CS-concept**

- **MD-concept**

- **doc-to-concept**

- **similarity matrix**

- **retrieval**

- **data**

- **brain**
SVD - Example

• \( \mathbf{A} = \mathbf{U} \Sigma \mathbf{V}^T \) - example:

\[
\begin{bmatrix}
1 & 1 & 1 & 0 & 0 \\
2 & 2 & 2 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 \\
5 & 5 & 5 & 0 & 0 \\
0 & 0 & 0 & 2 & 2 \\
0 & 0 & 0 & 3 & 3 \\
0 & 0 & 0 & 1 & 1 \\
\end{bmatrix}
\begin{bmatrix}
0.18 & 0 & 0.36 & 0 & 0.18 & 0 & 0.90 & 0 & 0.53 \\
0.53 & 0 & 0.80 & 0 & 0.27 & 0.58 & 0.58 & 0.58 & 0.71 & 0.71 \\
\end{bmatrix}
\begin{bmatrix}
9.64 & 0 & 0 & 5.29 \\
\end{bmatrix}
\]

SVD - Example

• \( \mathbf{A} = \mathbf{U} \Sigma \mathbf{V}^T \) - example:

\[
\begin{bmatrix}
1 & 1 & 1 & 0 & 0 \\
2 & 2 & 2 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 \\
5 & 5 & 5 & 0 & 0 \\
0 & 0 & 0 & 2 & 2 \\
0 & 0 & 0 & 3 & 3 \\
0 & 0 & 0 & 1 & 1 \\
\end{bmatrix}
\begin{bmatrix}
0.18 & 0 & 0.36 & 0 & 0.18 & 0 & 0.90 & 0 & 0.53 \\
0.53 & 0 & 0.80 & 0 & 0.27 & 0.58 & 0.58 & 0.58 & 0.71 & 0.71 \\
\end{bmatrix}
\begin{bmatrix}
9.64 & 0 & 0 & 5.29 \\
\end{bmatrix}
\]

SVD - Example

• \( \mathbf{A} = \mathbf{U} \Sigma \mathbf{V}^T \) - example:

\[
\begin{bmatrix}
1 & 1 & 1 & 0 & 0 \\
2 & 2 & 2 & 0 & 0 \\
1 & 1 & 1 & 0 & 0 \\
5 & 5 & 5 & 0 & 0 \\
0 & 0 & 0 & 2 & 2 \\
0 & 0 & 0 & 3 & 3 \\
0 & 0 & 0 & 1 & 1 \\
\end{bmatrix}
\begin{bmatrix}
0.18 & 0 & 0.36 & 0 & 0.18 & 0 & 0.90 & 0 & 0.53 \\
0.53 & 0 & 0.80 & 0 & 0.27 & 0.58 & 0.58 & 0.58 & 0.71 & 0.71 \\
\end{bmatrix}
\begin{bmatrix}
9.64 & 0 & 0 & 5.29 \\
\end{bmatrix}
\]
PCA interpretation

- best axis to project on: (‘best’ = min sum of squares of projection errors)

**Term2 (lung)**

**Term1 (data)**

PCA - interpretation

- PCA projects points onto the “best” axis
- minimum RMS error

**Term2 (retrieval)**

- first singular vector

**Term1 (data)**

Goal: extension to >=3 modes

\[ \mathbf{X} = [\mathbf{A}, \mathbf{B}, \mathbf{C}] = \sum_{r} \lambda_r \mathbf{a}_r \otimes \mathbf{b}_r \otimes \mathbf{c}_r \]
Specially Structured Tensors

- Tucker Tensor
- Kruskal Tensor

Our Notation

End of crash course

Bibliographic data (DBLP)

- Papers from VLDB and KDD conferences
- Construct 2nd order tensors with yearly windows with <author, keywords>
- Each tensor: 4584×3741
- 11 timestamps (years)
Multiway LSI

Authors:
- Michael Stonebraker, Michael Cherniack, Michael Carey, Hector Garcia-Molina, Jiawei Han, Jian Pei, Philip S. Yu, Jianyong Wang, Charu C. Aggarwal, Surajit Chaudhuri, Mitch Cherniack, Michael Stonebraker, H. Jagadish, Hector Garcia-Molina

Keywords:
- Tensor-based methods:
  - spot patterns and anomalies on time evolving graphs, and
  - on streams (monitoring)

Conclusions

Tensor-based methods:
- Two groups are correctly identified: Databases and Data mining
- People and concepts are drifting over time

Outline

- Problem definition / Motivation
- Static & dynamic laws; generators
- Tools: CenterPiece graphs; Tensors
- Other tools (Virus propagation, e-bay fraud detection)
- Conclusions
Virus propagation

- How do viruses/rumors/blog-influence propagate?
- Will a flu-like virus linger, or will it become extinct soon?

The model: SIS

- ‘Flu’ like: Susceptible-Infected-Susceptible
- Virus ‘strength’ $s = \frac{\beta}{\delta}$

Epidemic threshold $\tau$

of a graph: the value of $\tau$, such that

If $\text{strength } s = \frac{\beta}{\delta} < \tau$

an epidemic can not happen

Thus,
- given a graph
- compute its epidemic threshold
Epidemic threshold \( \tau \)

What should \( \tau \) depend on?

- avg. degree? and/or highest degree?
- and/or variance of degree?
- and/or third moment of degree?
- and/or diameter?

\[ \frac{\beta}{\delta} \leq \tau = \frac{1}{\lambda_{1, A}} \]

Proof: \([\text{Wang+03}]\)
### Outline

- Problem definition / Motivation
- Static & dynamic laws; generators
- Tools: CenterPiece graphs; Tensors
  - Other projects (Virus propagation, e-bay fraud detection)
- Conclusions

### E-bay Fraud detection

w/ Polo Chau & Shashank Pandit, CMU [WWW’07]
OVERALL CONCLUSIONS

- Graphs pose a wealth of fascinating problems
- self-similarity and power laws work, when textbook methods fail!
- New patterns (shrinking diameter!)
- New generator: Kronecker

References

References


References


References

Thank you!

- Christos Faloutsos
  www.cs.cmu.edu/~christos
  Wean Hall 7107

For more info on tensors:
www.cs.cmu.edu/~christos/TALKS/SIGMOD-07.pdf
3h version: www.cs.cmu.edu/~christos/TALKS/SDM-tut-07/