Data Mining – future directions, and past lessons

C. Faloutsos
CMU + Amazon (sabbatical)
Outline

• Credit where credit is due (12 foils)
• Future directions
• Past lessons: Listen
  – To the data
  – To domain-experts
• Conclusions
Thank you!

Prof. Ee-Peng Lim

Prof. Takashi Washio
Steering Committee

- Ee-Peng Lim
- P. Krishna Reddy
- Joshua Z. Huang
- Longbing Cao
- Jian Pei

- Myra Spiliopoulou
- Vincent S. Tseng
- Tru Hoang Cao
- Gill Dobbie
- Kyuseok Shim
GC and PC

- Geoff Webb
- Bao Ho
- Dinh Phung
- Vincent Tseng
Family

• Parents Nikos & Sophia
• Siblings Michalis, Petros, Maria
• Wife Christina
Academic ‘parent’

• Christodoulakis, Stavros (T.U.C.)
Academic ‘children’

• King-Ip (David) Lin
• Ibrahim Kamel
• Flip Korn
• Byoung-Kee Yi
• Leejay Wu
• Deepayan Chakrabarti
Academic ‘children’

- Jia-Yu (Tim) Pan
- Spiros Papadimitriou
- Jimeng Sun
- Jure Leskovec
- Hanghang Tong
Academic ‘children’

- Mary McGlohon
- Fan Guo
- Lei Li
- Leman Akoglu
- Dueng Horng (Polo) Chau
- Aditya Prakash
- U Kang
Academic ‘children’

- Danai Koutra
- Alex Beutel
- Vagelis Papalexakis
- Miguel Araujo
- Neil Shah
Academic ‘children’

- Bryan Hooi
- Hyun Ah Song
- Dhivya Eswaran
- Kijung Shin
- Namyong Park
Funding agencies/companies

- NSF (Maria Zemankova, Frank Olken, ++)
- DARPA, LLNL
- IBM, MS, HP, INTEL, Y!, Google, Symantec, Sony, Fujitsu, …
- Amazon
Outline

• Credit where credit is due
• Future directions
• Past lessons: Listen
  – To the data
  – To domain-experts
(Great time for Data Science)

- Alexa/Siri/Cortana
- Self-driving cars
- Alpha-go
- ...
Future directions:

- Time evolving graphs/networks
- What has a DBN learned?
- Explain the output
- Visualization
Future directions:

• Time evolving graphs/networks

• What has a DBN learned?

• Explain the output

• Visualization
Future directions:

• Time evolving graphs/networks
• What has a DBN learned?
• Explain the output
• Visualization
Future directions:

• Time evolving graphs/networks

• What has a DBN learned?
• Explain the output
• Visualization

• [how the brain works]
Outline

• Credit where credit is due
• Future directions
• Past lessons: Listen
  – To the data
    • D1: Clean data: a myth
    • D2: Surprises
  – To domain-experts
D1.1. Data & ‘cleanliness’

- Taxis
D1.1. Data & ‘cleanliness’

- Taxis
  - 0.1%: in the ocean
  - Longest taxi ride?
D1.1. Data & ‘cleanliness’

- Taxis
  - 0.1%: in the ocean
  - Longest taxi ride?
    - 6,000 miles
D1.2. Data & ‘cleanliness’

- Patients: ‘mode’ of age?

Rich Caruana
D1.2. Data & ‘cleanliness’

- Patients: ‘mode’ of age?
  - 99 (!)

Rich Caruana
D1.2. Data & ‘cleanliness’

• Patients: ‘mode’ of age?
  – 99 (!) and -99 (!!)

Rich Caruana
D1.2. Data & ‘cleanliness’

• Patients: ‘mode’ of age?
  - (99, or -99) for age

• Similarly, age of customer: -1
D1.2. Data & ‘cleanliness’

• Patients: ‘mode’ of age?
  - (99, or -99) for age

• Similarly, age of customer: -1
  - Fixing it -> $M in prediction accuracy
D1.3. Data & ‘cleanliness’

- Clicks, per hour of day
- NO periodicity

D1.3. Data & ‘cleanliness’

- Clicks, per hour of day
  - NO periodicity

- BUT: single user, 1 query/10sec
  - after removing him/her/it:
  - YES

Outline

• Credit where credit is due
• Future directions
• Past lessons: Listen
  – To the data
    • D1: Clean data: a myth
    • D2: Surprises
  – To domain-experts
D2.1 Growth of graph diameter

with Jure Leskovec (CMU -> Stanford)

and Jon Kleinberg (Cornell – sabb. @ CMU)

D2.1 Growth of graph 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?
D2.1 Growth of graph 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?
  - Diameter **shrinks** over time
D2.1. Diameter – “Patents”

- Patent citation network
- 25 years of data
- @1999
  - 2.9 M nodes
  - 16.5 M edges
D2.2. How many clusters?

- Eg.: clustering – k-means (or our favorite clustering algo)
- How many clusters are in the Sierpinski triangle?
D2.2. How many clusters?
D2.2. How many clusters?

K=3 clusters?
D2.2. How many clusters?

K=3 clusters?  
K=9 clusters?
D2.2. How many clusters?

• **Wrong question!** (‘How many line segments, to model a circle’)

![Diagram of a circle and a Sierpinski triangle.](image)
D2.2. How many clusters?

• But, does self-similarity appear in real life?
Outline

• Credit where credit is due
• Future directions
• Past lessons: Listen
  – To the data
  – To domain-experts
    • E1: fractals / self-similarity
    • E2: power-laws
B. Mandelbrot

The Fractal geometry of nature, 1982

2 pages of self-similar objects:
• Bark of trees
• Surface of mountains
• Human lungs
• Surface of mammalian brain

PAKDD'18 C. Faloutsos
E1.1. Real, self similar dataset
E1.1. Real, self similar dataset
E1.1. Real, self similar dataset
E1.1. Real, self similar dataset
• the red is true
• origin: Norway
• but most other coastlines are ‘self-similar’, too!
E1.2. Disk traffic

- disk traces: self-similar:
E1.3. Web traffic

- [Crovella, Bestavros, SIGMETRICS’96]

1000 sec; 100sec
10sec; 1sec
Outline

• Credit where credit is due
• Future directions
• Past lessons: Listen
  – To the data
  – To domain-experts
    • E1: fractals / self-similarity
    • E2: power-laws
Fractals <-> power laws, eg.:

\[ N(r) = r^{\log(3)/\log(2)} = r^{1.58} \]
Fractals <-> power laws, eg.:

\[ N(r) = r^{\log(3)/\log(2)} = r^{1.58} \]

\[ y = x^a \quad a = 1.58 \]
E2.1. : 2x mass -> 2x food?
E2.1. : 2x mass \neq 2x food?
Metabolic rate

Experts say:

http://universe-review.ca/R10-35-metabolic.htm
Experts say:

Metabolic rate

3/4

http://universe-review.ca/R10-35-metabolic.htm
Metabolic rate

Kleiber's law:

http://universe-review.ca/R10-35-metabolic.htm
E2.2.: Triangle Patterns

- Real social networks have a lot of triangles
E2.2.: Triangle Patterns

• Real social networks have a lot of triangles
  • Friends of friends are friends
• Any patterns?
  • 2x the friends, 2x the triangles?
E2.2.: Triangle Patterns
[Tsourakakis ICDM 2008]

Reuters

Epinions

X-axis: degree
Y-axis: mean # triangles
n friends $\rightarrow \sim n^{1.6}$ triangles

DTPL

Mean #Triangles

Degree

Mean #Triangles

Degree

DTPL

Mean #Triangles

Degree

DTPL

Mean #Triangles

Degree
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]
Summary

• Golden age of Data Science / Data Mining

• Data:
  – Never ‘clean’
  – Often: surprises

• Domain experts – **cross-disciplinarity:**
  – Help us avoid surprises
Parting joke:

- Data scientists spend 80% of their time, cleaning data;
Parting joke:

• Data scientists spend 80% of their time, cleaning data;
• And the rest 20% complaining about it.
Thank you!

Listen to experts -> Reach out

Listen to data