Big Data Provenance

Daniel Deutch

Database Group
Blavatnik School of Computer Science,
Raymond and Beverly Sackler Faculty of Exact Sciences
Tel Aviv University

"Your recent Amazon purchases, Tweet score and location history makes you 23.5% welcome here."

cartoon by T. Gregorius
Data Management

• Since the emergence of the Web, Data is everywhere
  – Databases, wiki, blogs, social data,…

• Many flavors and sources lead to huge, unreliable and heterogeneous Data

• Data management researchers aim at developing generic tools for finding, organizing, and analyzing Data

• Many branches for each high-level goal, and a lot of research in each branch
Information Extraction

Johannes Hoffart, Fabian Suchanek, Klaus Berberich, Gerhard Weikum

YAGO2: A Spatially and Temporally Enhanced Knowledge Base from Wikipedia. AI J. 2013
Crowdsourcing

Make Money by working on HITs

- HITs - Human Intelligence Tasks - are individual tasks that you work on. Find HITs now.
- As a Mechanical Turk Worker you:
  - Can work from home
  - Choose your own work hours
  - Get paid for doing good work

Get Results from Mechanical Turk Workers

- Ask workers to complete HITs - Human Intelligence Tasks - and get results using Mechanical Turk. Register Now
- As a Mechanical Turk Requester you:
  - Have access to a global, on-demand, 24 x 7 workforce
  - Get thousands of HITs completed in minutes
  - Pay only when you're satisfied with the results

www.mturk.com

Amsterdamer, Grossman, Milo, Senellart,
Crowd Mining, SIGMOD 2013
Web Applications

www.ebay.com

D., Moskovitch, Tannen, PROPOLIS: Provisioned Analysis of Data-Centric Processes. VLDB 2013
Scientific Workflows

(a) M1: Compute alignment → M2: Refine alignment

M7: Align sequences

M3: Iterate over seeds → M4: Find MP trees

M5: Check exit condition → M6: Compute consensus

M8: Infer trees

M9: Infer tree

(b) S1:M1 → S2:M2 → S3:M3 → S4:M4 → S5:M5

S6:M3 → S7:M4 → S8:M5 → S9:M6 → O

Alignment2, Tree1, Tree3, Seed1

(c) seq1, seq2, seq3, seq4, seq5, seq6, seq7, seq8, seq9, seq10

S1:M1 → S2:M2

Alignment1

Alignment2

Tree1, Tree2, Tree3, Tree4, Tree5, Tree6

Davidson et. al,
Provenance in Workflows Systems,
DEBU 2007
Common Problem

- Data-intensive systems are highly complex
  - Manipulate big-scale data in intricate ways

- Error-prone
  - Errors in input (measurements, crowd, text)
  - Errors in processing (ambiguities, imperfect text understanding, “bugs”)

- Precision vs. recall tradeoff

- Lots of domain and application-specific efforts on improving both simplicity and precision
Provenance

Art provenance

Wine provenance
Data Provenance

• Tracking **where** Data came from,
  how it was **extracted**
  how it was **manipulated**

• We build generic algorithms and tools for provenance **tracking, presentation, and analysis**

• Assisting in **error detection, debugging** and better **understanding** of the application

• Fundamental challenge: Big Data, even **bigger provenance**
D., Gilad, Moskovitch,
**selP: Selective Tracking and Presentation of Data Provenance,**
to appear in ICDE ‘15
D., Moskovitch, Tannen, 
PROPOLIS: Provisioned Analysis of Data-Centric Processes, VLDB ‘13
A Provenance Framework for Data-Dependent Process Analysis, VLDB ‘14
Approximate Provenance for Crowdsourcing

• Instead of computing “exact” (partial/compact) provenance we may also approximate

• Useful in crowdsourcing where multiple answers are aggregated, exact individual sources less important to track
  – Definition of approximation depends on intended use

• Instrument crowdsourcing systems so that they track approximate provenance

Ainy, Davidson, D., Milo, *Approximated Provenance for Complex Applications*, Tapp 2014
Explaining computation

• Provenance tracking leads to better applications and more reliable data

• Seamless provenance tracking through tools
  – Allow application owners to easily integrate provenance solutions
  – With reasonable overhead in time and storage

• Privacy issues
  – When provenance is shown to external user
  – Partially addressed through selective tracking
"Your recent Amazon purchases, Tweet score and location history makes you 23.5% welcome here."

cartoon by T. Gregorius