GRAPH DB’S & APPLICATIONS

Denis Vrdoljak  |  Gunnar Kleemann

UC Berkeley School of Information
Berkeley Data Science Group, LLC
PRESENTATION ROAD MAP

- Intro
- Background
- Examples
- Our Work
- Graph Databases
Intro

Background
Examples
Our Work
Graph Databases
About Us: BDSG

Berkeley Data Science Group

Founded by UC Berkeley Data Science instructors and alumni with the goal of bringing Berkeley data science projects to market and commercializing Berkeley Data Science research.
About Us: the Speakers

Denis Vrdoljak
denis@bds.group
dvrdolja@cisco.com

Gunnar Kleemann, PhD
gunnar@bds.group
gunnarkl@berkeley.edu
Why Graphs?

Why Graph Databases?
GRAPH DB’S OPTIMIZED FOR RELATIONSHIPS

- **Graph Databases** store data in tables/rows/columns, just like a traditional RDBMS
- **First Class Citizen** is a relationship, not an entity
- **Graph DB’s** are optimized for graph traversals
- This also makes them slow at data retrieval
- **But, they’re a LOT faster at traversing the nodes of a graph!**
Intro
Background
Examples
Our Work
Graph Databases
WHAT IS GRAPH THEORY?

DATES BACK TO 1736: SEVEN BRIDGES OF KÖNIGSBERG, BY LEONHARD EULER

LAID DOWN THE ORIGINAL GROUNDWORK FOR WHAT BECAME GRAPH THEORY

ALSO EVOLVED INTO MODERN DAY NETWORK ANALYSIS (OR NETWORK GRAPH ANALYSIS) AND SOCIAL NETWORK ANALYSIS (SNA)
THE SEVEN BRIDGES OF KONIGSBERG

**Problem Statement:** Can you visit each part of the town, using each bridge only once?

**Leonhard Euler** came up with a new way of thinking about the problem, and in turn became the father of modern **Graph Theory**.
RECOMMENDER SYSTEMS: TRADITIONAL VS. GRAPH

**Traditional:**
- At Scale
- Production Deployments

**Graph-Based:**
- Exploratory
- Highly Contextual
- Known Rules
APPLICATIONS OF GRAPH THEORY

- **Social network analysis**
- **Map / GPS algorithms - shortest distance between two points, etc.**
- **AI algorithms**
- **Search engine algorithms**
EXAMPLES OF GRAPH APPLICATIONS

- 9/11 TERRORIST NETWORK
- LONDON PHONE NETWORK
- ENRON EMAILS
- PANAMA PAPERS
Intro
Background
Examples
Our Work
Graph Databases
Keyword Rec: an HR Keyword Assistant
CONSIDER A TYPICAL JOB REQUISITION

Job Title
- topic information

Company and institute
- workplace
- geographic location

Keywords
- skills
- experience/focus
LOOKING AT A SKILL CAN SUGGEST A GEOGRAPHIC HUB

City

San Francisco

Austin

Company

SanDisk

Pfizer

Genentech

Asuragen

Job Title

Bioinformatics

Skill

DNA analysis
LOOKING AT A REGION WE MIGHT SUGGEST CRUCIAL SKILLS

- **City**: Austin
- **Company**: Asuragen, Rackspace
- **Job Title**: Data scientist, Bioinformatics, Database manager
- **Skill**: R, Python, DNA analysis, SQL
WE CAN SUGGEST DEFICIENT SKILLS
relatedskills('Python', 'SV', 5)

[['python', 328],
 ('similar', 137),
 ('unix', 136),
 ('programming experience', 21),
 ('language', 15)]

relatedskills('Python', 'NY', 5)

[['plus', 636],
 ('programming experience', 615),
 ('python', 383),
 ('sql', 324),
 ('r', 320)]
SF vs NY: Who has better data scientists?

https://public.tableau.com/profile/denis.vrdoljak#!/vizhome/SVvsNY_JobSkillsPercentCompare/Sheet3?publish=yes
## Who has Better Data Scientists?

### Sheet 1

<table>
<thead>
<tr>
<th>Skill</th>
<th>Experience</th>
<th>Financial</th>
<th>Full</th>
<th>Google</th>
<th>Health</th>
<th>Job</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Region</th>
<th>New York</th>
<th>Silicon Valley</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

The chart compares the skill levels in different regions, with New York and Silicon Valley being the two regions of interest.
Who has Better Data Scientists?

Sheet 1

<table>
<thead>
<tr>
<th>Skill</th>
<th>New York</th>
<th>Silicon Valley</th>
</tr>
</thead>
<tbody>
<tr>
<td>market</td>
<td></td>
<td></td>
</tr>
<tr>
<td>marketing</td>
<td></td>
<td></td>
</tr>
<tr>
<td>office</td>
<td></td>
<td></td>
</tr>
<tr>
<td>plan</td>
<td></td>
<td></td>
</tr>
<tr>
<td>position</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Region
- Blue: New York
- Orange: Silicon Valley
Who has Better Data Scientists?

Sheet 1

<table>
<thead>
<tr>
<th>Skill</th>
<th>New York</th>
<th>Silicon Valley</th>
</tr>
</thead>
<tbody>
<tr>
<td>database</td>
<td></td>
<td></td>
</tr>
<tr>
<td>degree</td>
<td></td>
<td></td>
</tr>
<tr>
<td>design</td>
<td></td>
<td></td>
</tr>
<tr>
<td>development</td>
<td></td>
<td></td>
</tr>
<tr>
<td>engineering</td>
<td></td>
<td></td>
</tr>
<tr>
<td>environment</td>
<td></td>
<td></td>
</tr>
<tr>
<td>experience</td>
<td></td>
<td></td>
</tr>
<tr>
<td>financial</td>
<td></td>
<td></td>
</tr>
<tr>
<td>full</td>
<td></td>
<td></td>
</tr>
<tr>
<td>google</td>
<td></td>
<td></td>
</tr>
<tr>
<td>health</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Who has Better Data Scientists?
...now let’s take a look back to our very first project, and what we’ve done with it since!
BioRevs: Predicting Biotech IPO Rates through Collaboration Networks
CONDUCTING A JOB SEARCH IN BIOTECH

YOU COULD IDENTIFY BIOTECHNOLOGY HUBS BY COUNTING COMPANIES
BIOTECH HUBS CAN BE PROFILED BY IPO AND % SCIENTISTS

Percent IPO = % Companies reaching IPO in < 6000 days

National average is ~16% reach IPO.

National average is ~21% Scientists.
BASED ON TRADITIONAL MACHINE LEARNING ANALYSIS
WE BUILT BIOREVS

* 23 and me

* Data up to 2013

* Theranos
WE CAN USE A GRAPH TO GET MORE
PUBLICATIONS CAN BE PARSED INTO IMPORTANT DATA

**Journal -**
- Discipline
- Audience
- Impact

**Title -**
- topic information

**Collaborator list**
- professional relationships

**Company and institute**
- workplace
- geographic location
PUBMED DATABASE
THE WORLD’S BIOMEDICAL RESEARCH

24,000,000

FULL PUBMED DATASET (ALL BIOMEDICAL LITERATURE)

~ 1.5 MILLION OPEN SOURCE

PUBMED CENTRAL (BIOTECH-OPEN ACCESS SUBSET)

DATA INCLUDE: PUBLICATION TITLES, AUTHORS, AND OTHER METADATA
PUBLICATION AND COMPANY DATA IN A GRAPH

City:
- San Francisco
- Austin

Institution:
- Pfizer
- Genentech
- Asuragen

Author:
- Melissa
- Gunnar
- Fred

Paper:
- id44234
- id12234

Title words:
- stress
- cancer
- pathogen
CITIES DIFFER IN SCIENTIFIC EXPERTISE

San Diego

Austin

San Francisco
QUANTIFICATION OF SCIENCE NETWORKS WITH COLLABORATION GRAPHS
WE CAN GET THE COLLABORATION NETWORK TOPOLOGY FROM THE GRAPH
SOME TYPICAL PUBLICATION PATTERNS

SCOTT EMMONS’
2ND DEGREE
COLLABORATION
NETWORK

COLEEN MURPHY’S
2ND DEGREE
COLLABORATION
NETWORK
INCREASING NODE COUNT WITH DISTANCE FROM AUTHOR

PAPERS < COAUTHORS < SECONDARY < TERTIARY
WE CAN SEE AN OUTLIER!

PAPERS <
COAUTHORS <
SECONDARY <
TERTIARY
A BIG PUBLISHER WITH 2055 LINKS!

“ONE OF THE MOST PRODUCTIVE RESEARCHERS, EDITORS, AND PUBLISHERS IN THE ONLINE HEALTH FIELD.”

IN 2004 RECEIVED THE JANSSEN-CILAG FUTURE AWARD, REFERRED TO AS THE GERMAN “HEALTH CARE NOBEL PRIZE”.

FOUNDER OF AN ACADEMIC FIELD!
ASSOCIATION BETWEEN SEARCH ENGINE QUERIES AND INFLUENZA INCIDENCE,
HE COINED THE TERMS "INFOVEILLANCE" AND "INFODEMIIOLOGY" FOR THESE KINDS OF APPROACHES.
FEW CO-AUTHORS AND MANY PAPERS SUSPICIOUS PATTERN

But he is only cited for 120 papers and 40 book chapters.

So what are those other 1900 links?

Probably editing jobs (false positives)
Intro
Background
Examples
Our Work

Graph Databases
SOME KEY GRAPH DATABASES

- Neo4J (well supported)
- Titan/Janus Graph (distributed backend)
- AensGraph (Postgres compatible)
- Grakn (knowledge graph)
NEO4J

- **Industry Standard Features**
- **Large Userbase and Developer Community**
- **Built from the ground up as a graph database**

https://neo4j.com/
TITAN/JANUS GRAPH

- Apache project
- Early adopter of distributed backend
- Elastic scalability
- Integration with Tinkerpop graph stack
- Multiple user access
- Real time updates

https://www.predictiveanalyticstoday.com/titan/
AGENSGRAPH

- Highly performant graph database
- Hybrid database built on PostgreSQL
- SQL and Cypher in the same query

http://bitnine.net/agensgraph/?ck=1
GRAKN.AI

- **Knowledge representation in graphs for AI purposes**
  - Nodes represent “objects”, and edges are relationships between them.
- **SQL-type query language, GraQL**, used to quickly and intuitively make queries in the knowledge graph
- **Steadily growing technology**

https://grakn.ai/
Thank You!

• Denis Vrdoljak, MIDS
  Managing Director, BDSG
  Marketing Analyst/Data Scientist, Cisco

• denis@bds.group

• dvrdolja@cisco.com

Gunnar Kleemann, PhD, MIDS
Senior Data Scientist, BDSG
Data Science Instructor, UC Berkeley

• gunnar@bds.group

• gunnarkl@berkeley.edu