
ORACLE®

What Are They Talking About These Days?

Analyzing Topics with Graphs

Korbinian Schmid
<korbi.schmid@oracle.com>
Oracle Spatial&Graph, Austin, USA

Charles Wang
<charles.wang@oracle.com>
Oracle Spatial&Graph, Beijing, China

Davide Basilio Bartolini
<davide.bartolini@oracle.com>
Oracle Labs, Zürich, Switzerland

Damien Hilloulin
<damien.hilloulin@oracle.com>
Oracle Labs, Zürich, Switzerland

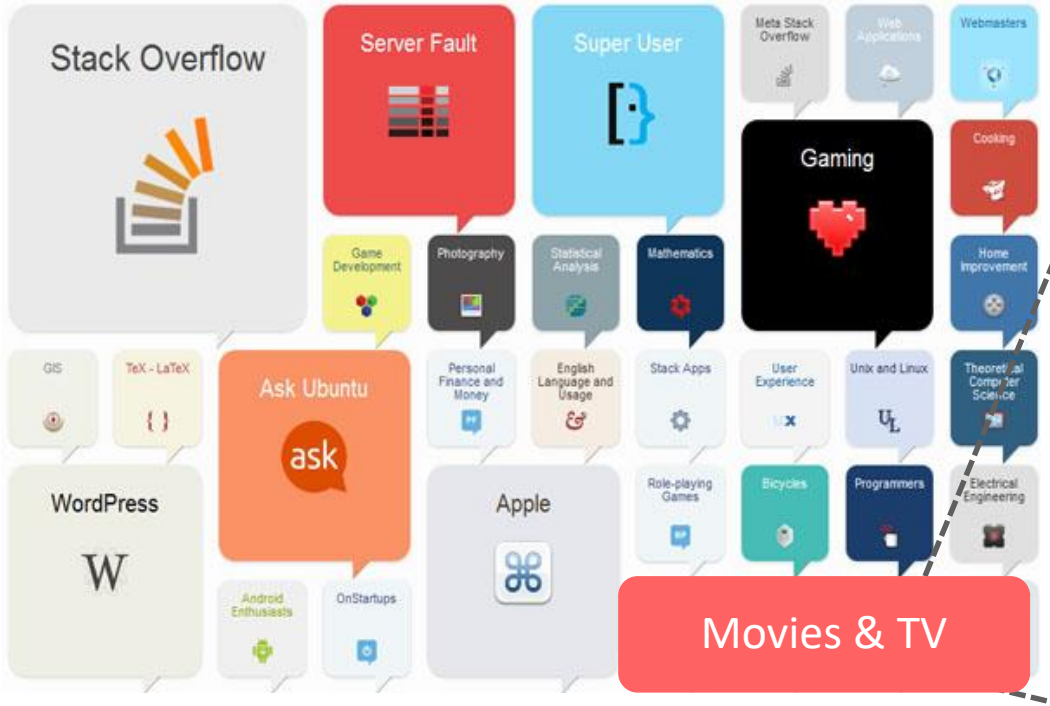
Safe Harbor Statement

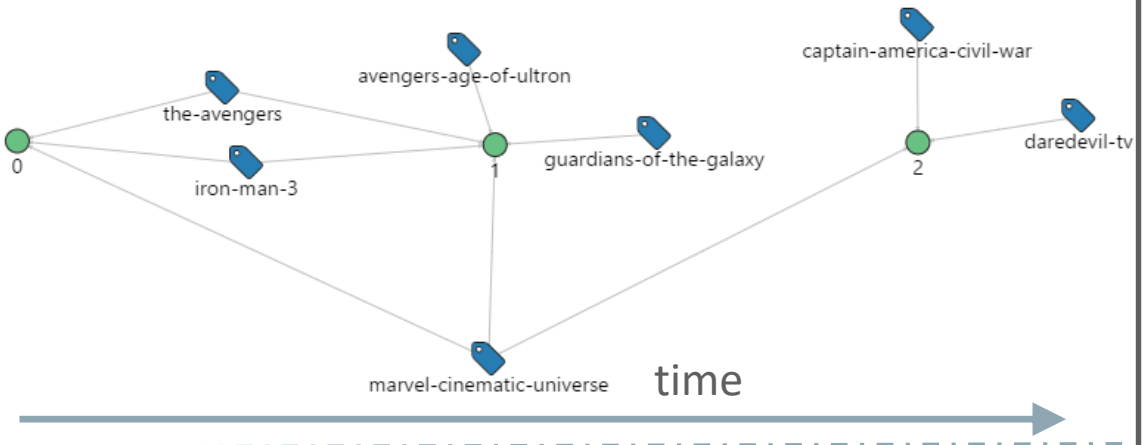
The following is intended to outline our general product direction. It is intended for information purposes only, and may not be incorporated into any contract. It is not a commitment to deliver any material, code, or functionality, and should not be relied upon in making purchasing decisions. The development, release, and timing of any features or functionality described for Oracle's products remains at the sole discretion of Oracle.

In a Nutshell

What are the topics people are discussing and how do they evolve?

Who are the experts in a given topic?





username	numberOfPosts	commmunity
Andrew Martin	5	6202
Ankit Sharma	5	6202
Walt	4	6202
BCdotWEB	4	6202
Nobby	3	6202

Session Agenda

1. Dataset and Objectives
2. Lightning Primer on Topic Modeling
3. Looking at this as a Graph
4. Tools
5. Demo time

Dataset and Objectives

Dataset info

- Anonymized dump of all user-contributed content on the Stack Exchange network [1]
- The data contains posts, comments, user profiles, tags, ...
- In the demo, will look at the **movies & tv** site from the beginning (2011) until March 2017

Objectives

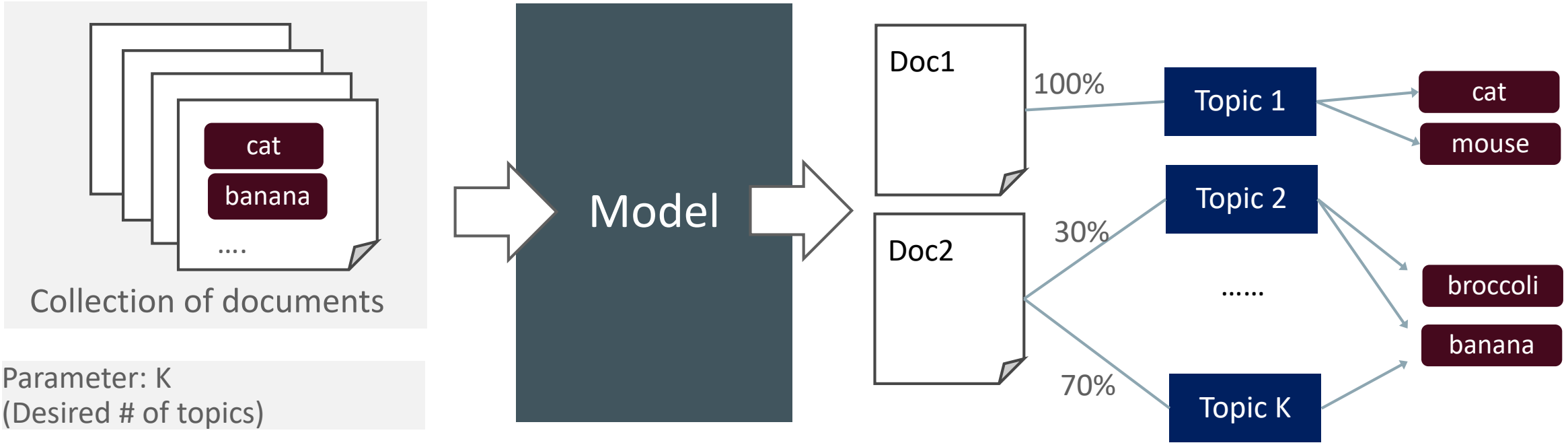
1. Topic modeling
   - Identify broader discussion topics from tags and posts
   - Analyze how topics change over time (topic evolution) and their importance (topic strength)
2. User analysis
   - Find who are the experts in a given topic

[1] The data is available at https://archive.org/details/stackexchange

Lightning Primer on Topic Modeling

"A topic model is a type of statistical model for discovering the abstract "topics" that occur in a collection of documents." [1]

Classic approach (e.g., LDA) quick overview:



Parameter: K
(Desired # of topics)

[1] https://en.wikipedia.org/wiki/Topic_model

Our dataset has more structure than a doc. collection

Stackexchange dump original data in xml format

- Posts.xml

```
<row Id="1" AcceptedAnswerId="42" CreationDate="…" Body="…" OwnerUserId="37"
Title="…" Tags="t1;t2;…" />
```

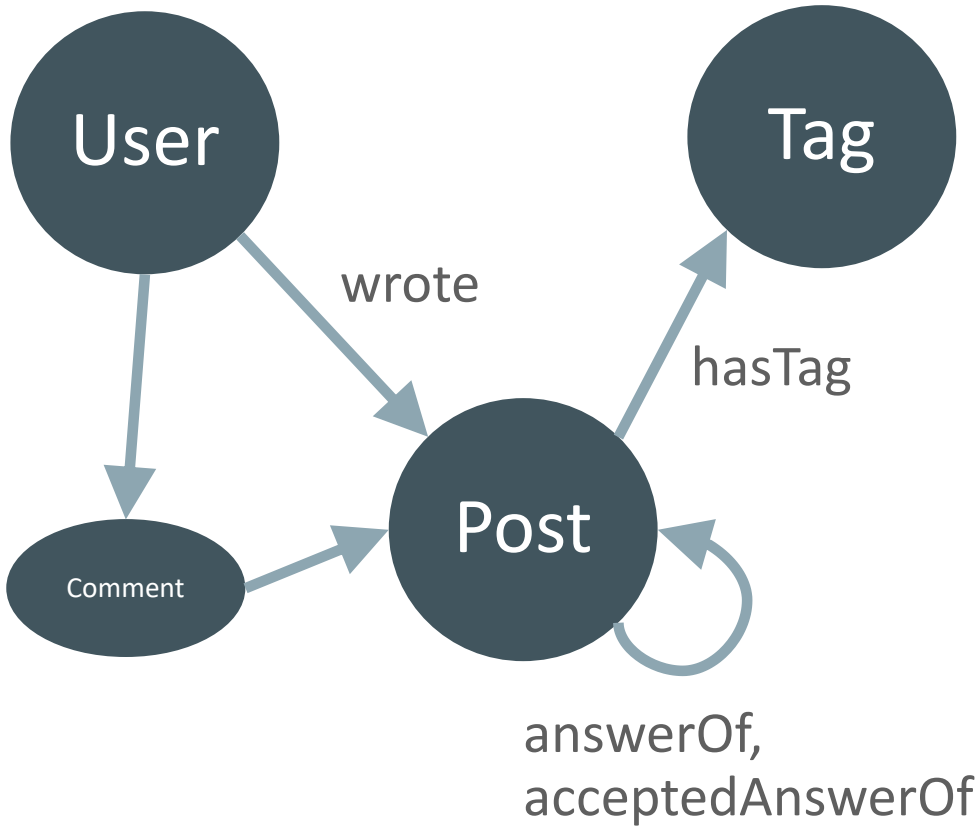
- Tags.xml

```
<row Id="1" TagName="harry-potter" Count="2" />
```

- Users.xml

```
<row Id="15" DisplayName="..." />
```

Graph for This Demo



- Preprocessed the data to create the property graph structure from the xml files (won't focus on this step in this demo)
- The result is a graph in "EDGE_LIST" format, that we can load into PGX [1]
- Also add "reverse edges" for Tag <-> Post and Post <-> Post, to highlight community structure

[1] https://docs.oracle.com/cd/E56133_01/latest/reference/loader/index.html

Topic Analysis Approach

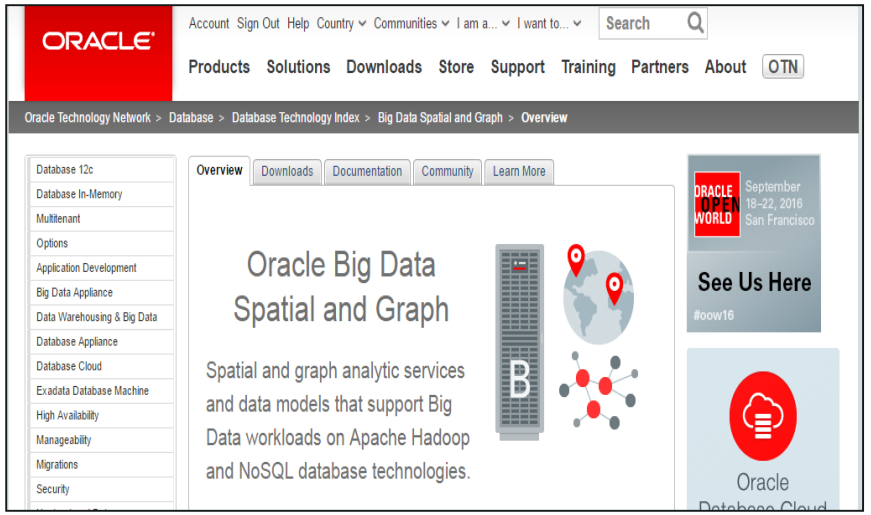
- Find communities of vertices in a graph via the “map equation” [1]
  - Metric: minimize entropy in vertex labeling using communities as prefixes
  - Using Relaxmap [2], a parallel version of the algorithm
- Each community can be interpreted as a topic
  - We find “clear-cut” topics, not a probabilistic mixture
  - Posts / tags are assigned to one and only one topic

[1] http://www.mapequation.org/ [2] http://uwescience.github.io/RelaxMap/

Oracle PGX

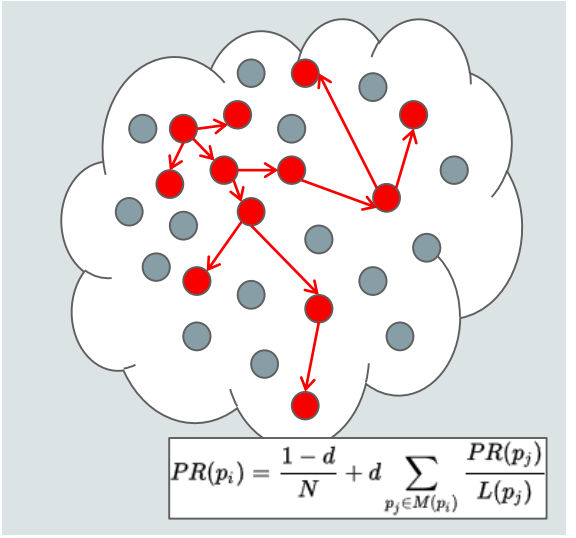
A graph analysis package developed at Oracle Labs

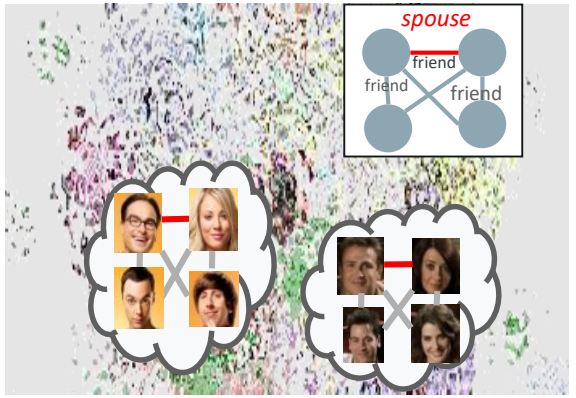
- Fast and easy application of graph analysis to the dataset
- Already integrated with Oracle products
  - Big Data Spatial and Graph (BDSG)
  - Spatial and Graph (RDBMS 12.2c)
  - Advanced Analytics (OAA) – upcoming



Property Graph Analysis Strategies

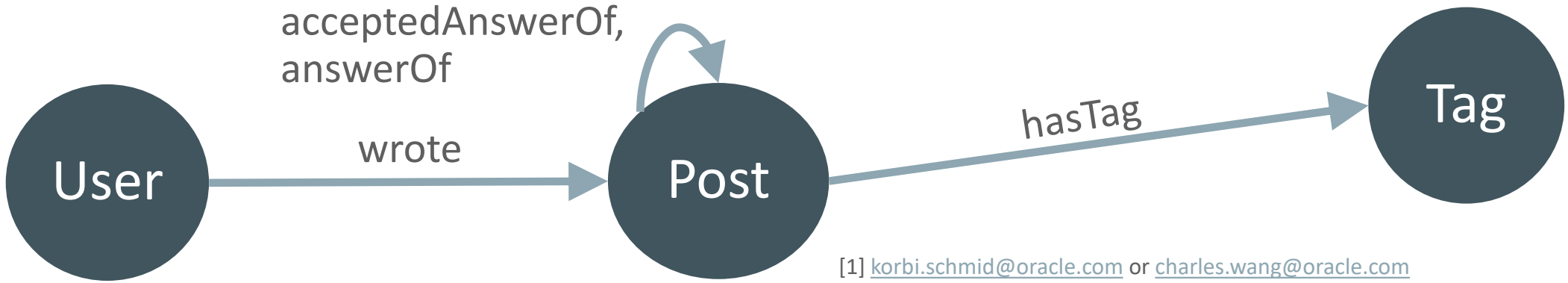
- Two major approaches
  - Computational graph analytics
    - Iterate the graph and compute properties / stats
  - Graph pattern matching
    - Query the graph to find sub-graphs that match a pattern
- PGX supports both approaches
  - Green Marl / Java API
  - PGQL
- And offers an environment to mix them
  - Groovy shell, OL Data Studio (under development)





Demo Time!

- The demo is based on the Oracle Labs Data Studio frontend
  - A notebook-like environment supporting %pgx and %pgql interpreters (and others)
  - Not yet released (but it's "just a frontend", can do the same analysis with PGX alone)
  - We are planning to release this demo in an upcoming tech preview
  - Contact us [1] or poll the PGX OTN website [2] for updates
- Demo focuses on the "movies & tv" forum from stackexchange



[1] korbi.schmid@oracle.com or charles.wang@oracle.com
[2] http://www.oracle.com/technetwork/oracle-labs/parallel-graph-analytics/overview/index.html

Integrated Cloud

Applications & Platform Services


ORACLE®