[DEV5420] When Graphs Meet Machine Learning

Rhicheek Patra
Senior Member of Technical Staff
Oracle Labs

Sungpack Hong
Research Director
Oracle Labs

Jinha Kim
Principal Member of Technical Staff
Oracle Labs

Live for the Code

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, timing, and pricing of any features or functionality described for Oracle's products may change and remains at the sole discretion of Oracle Corporation.

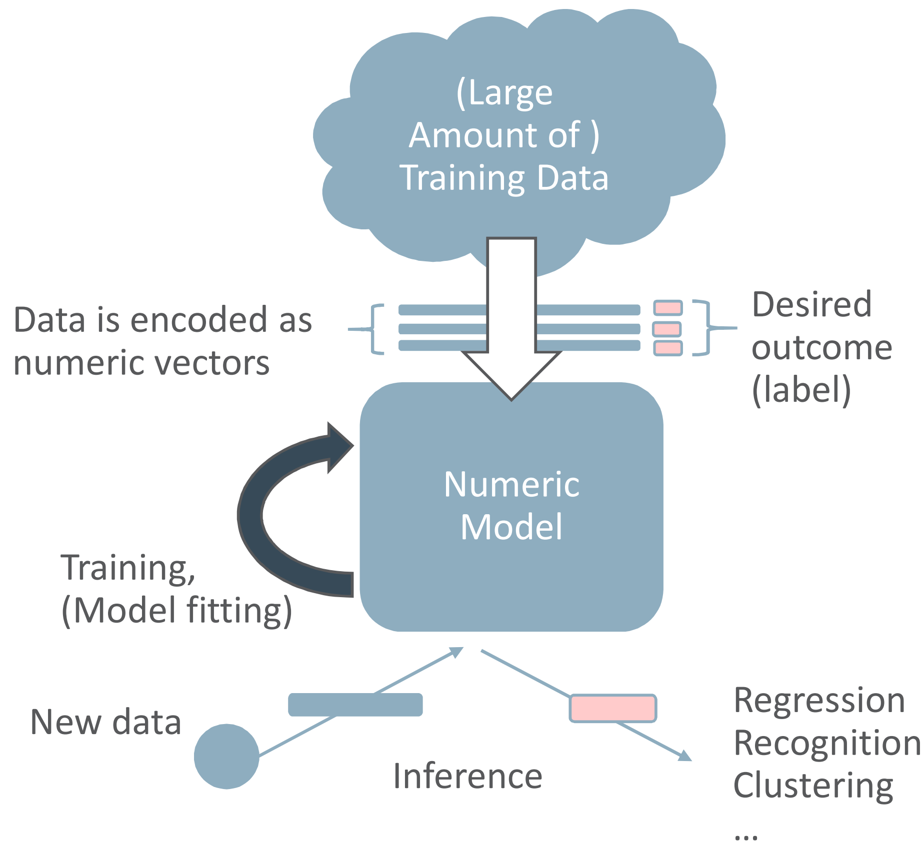
Agenda

1. Machine Learning and Graph Analysis
2. Encoding Relationship Distance
3. Encoding Irregular Structure
4. Current Works and Directions

ORACLE

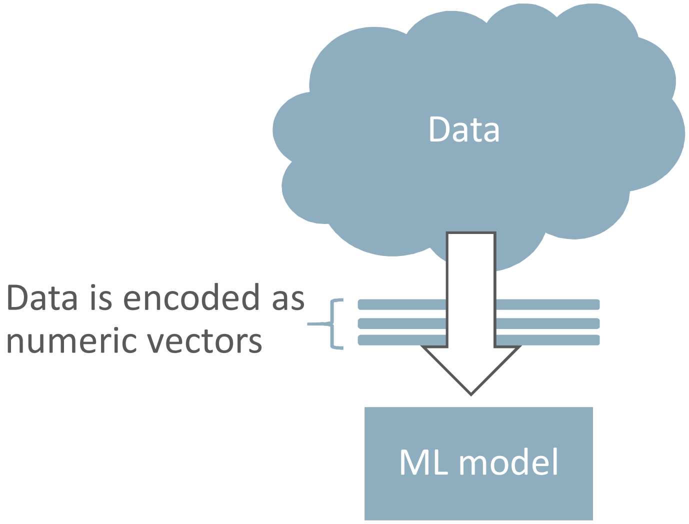
Machine Learning

- A very popular concept these days.
  - A system that progressively improves its performance from data
- How it's done: Supervised Learning*
  - Training Data, encoded as numeric vectors
  - Feed into numeric model
  - Train or fit the model to produce desired outcome for given data
  - Given new data, the model can infer its outcome
  - (for tasks like regression, recognition, clustering …)



What’s the problem, then?

- What is your data?
  - More precisely, what is the information that you want to exploit from your data?
  - How that information can be encoded (as numeric vectors) to feed into ML model?



Example) Personalized Product Recommendation

Would like to recommend items
that are purchased a lot by *similar* people
who had purchased a lot of *similar* items
that are purchased by the target customer
(i.e. popular to people-like-me)

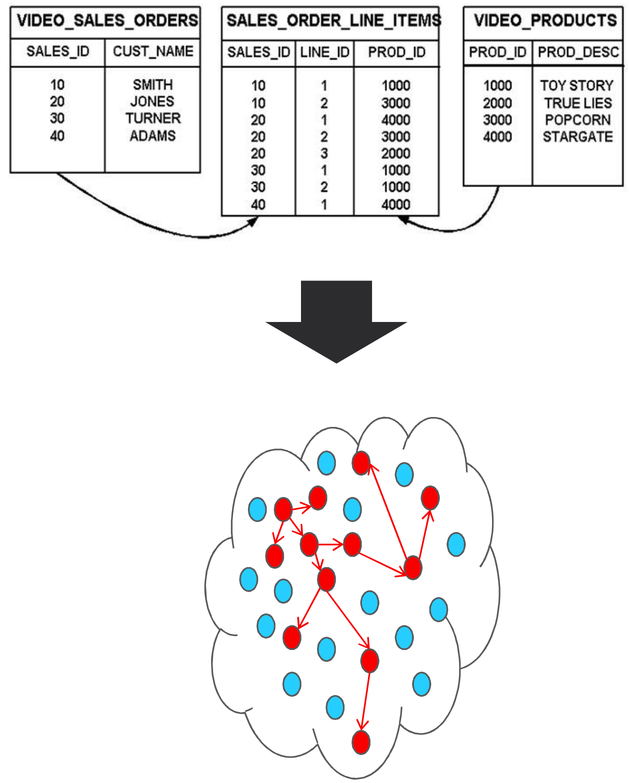
How to extract this information?
How to encode this information and feed into ML pipeline?

Graph Modeling and Analysis

- Graph Modeling
  - Represent your (relational) dataset as a graph
  - Entities become vertices
  - Relationships become edges

➔ Fined-grained relationships are captured in graph

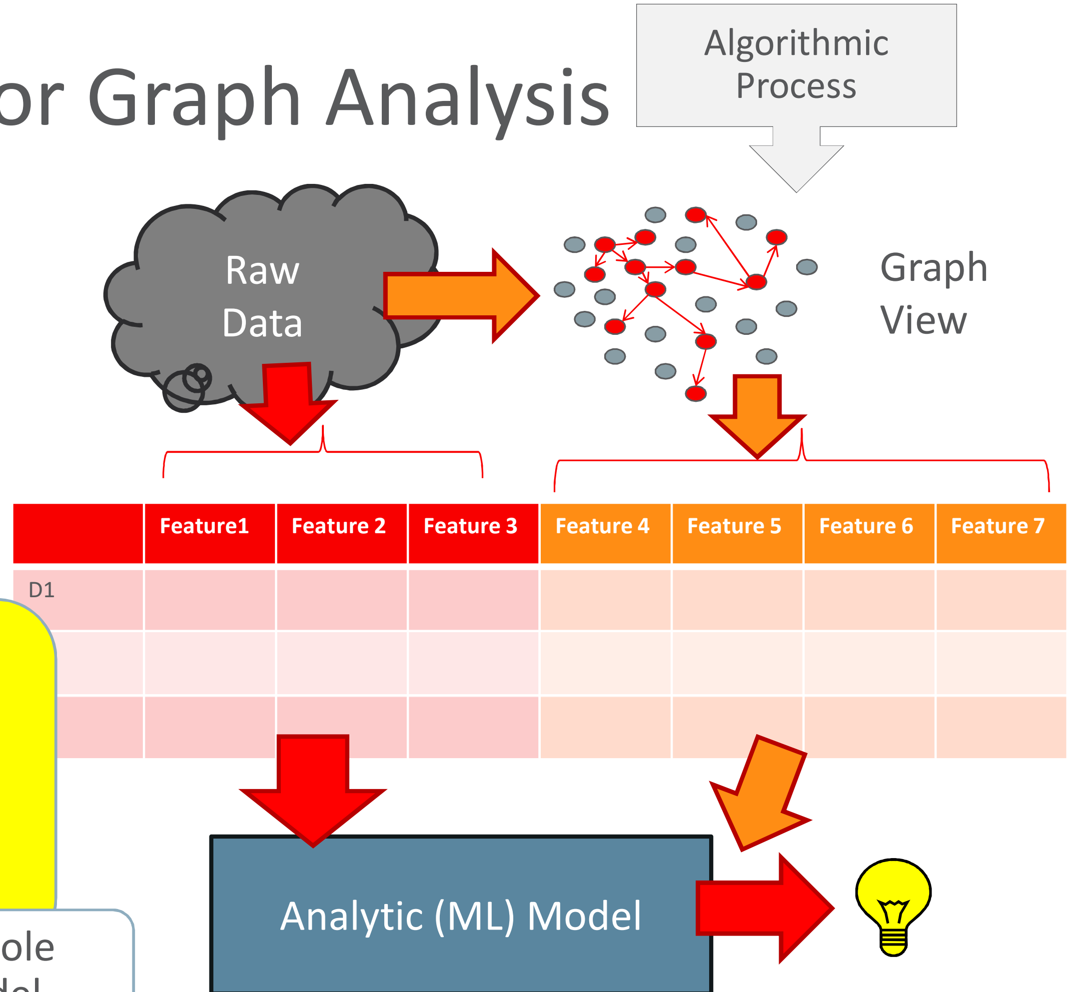
- Graph Benefits
  - Quickly query multi-hop relationships
  - Visualize your data and explore it interactively
  - Analyze your data using graph signals ✔



ORACLE

Conventional Approach for Graph Analysis

- Graph representation captures fine-grained relationship between data entities
- By applying *graph algorithms*, one can get useful information from the graph
- Or produce additional features that can be added into analytic (ML) model



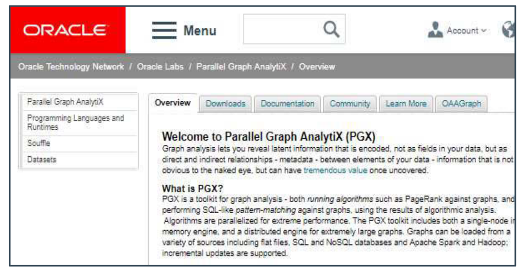
- More conventional , algorithmic approach
- Still ***very*** effective. Requires no training data (free from cold starting problem)
- Will cover more in another session: *[DEV5397] Automate Anomaly Detection with Graph Analytics*

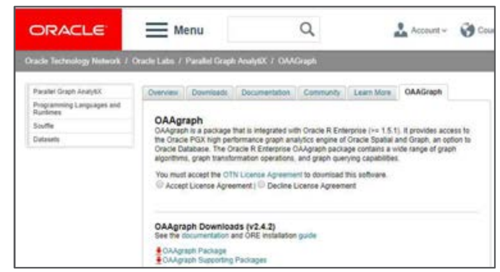
➔ This talk tries to feed the whole graph information into ML model

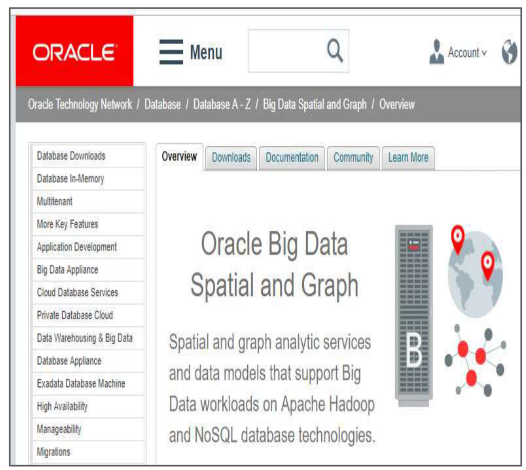
ORACLE

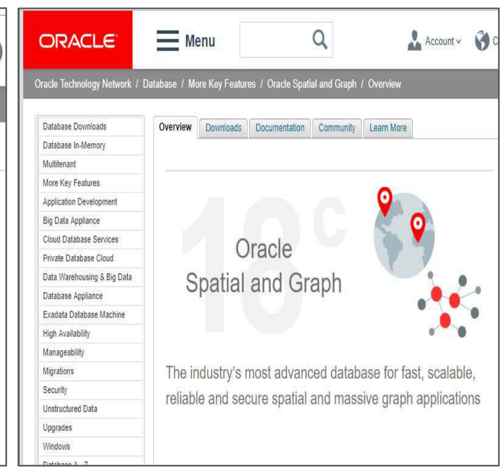
Graph Tooling in Oracle

- (Not going too deep here)
- Oracle provides graph technology with several different flavors
  - Oracle Spatial and Graph:Database
  - BDSG: Big Data Appliance
  - OAA.Graph (R): Advanced Analytics
  - Graph Cloud Service
- Key components are shared between these projects









ORACLE

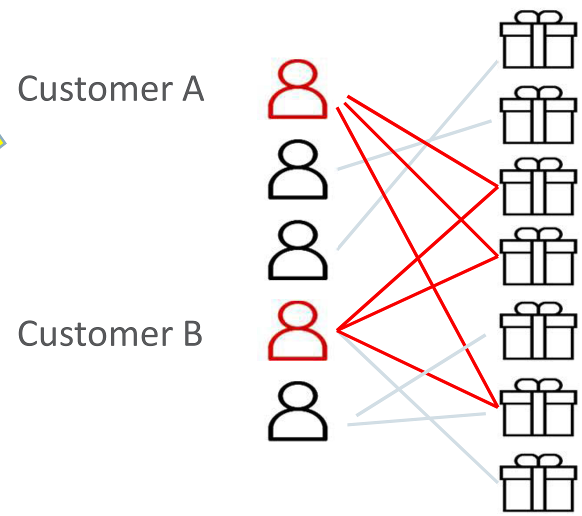
Agenda

1. Machine Learning and Graph Analysis
2. Encoding Relationship Distance
3. Encoding Irregular Structure
4. Current Works and Directions

Data Entity Distance in Graph

- Graph captures fine-grained relationship between data entities
  ➔ Closeness by such relationship can be defined and measured on the graph

Customer A and B are *close* to each other
(because they purchased the same items a lot)



Graph gives you several ways to define and measure distance between vertices:

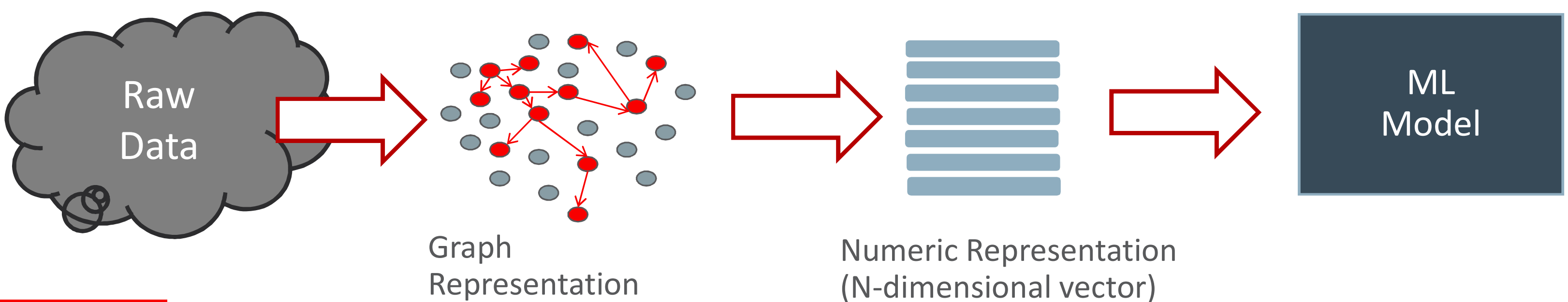
- Shortest path
- Hop distance
- Maximum flow
- Common neighbor count
- Random walk distance
- …

Going Back to the Problem

- Graph captures distance between data entities.
- You want feed this distance information into your ML pipeline
- You need numeric representation of your data that retains the distance information

x, y: data entity (represented as vertex in graph)
v(x), v(y): n-dimensionsal vector representation of x and y

x, y close in graph ➔ || *v(x)* - *v(y)* || small in n-dimensional vector space



ORACLE

How to achieve this?

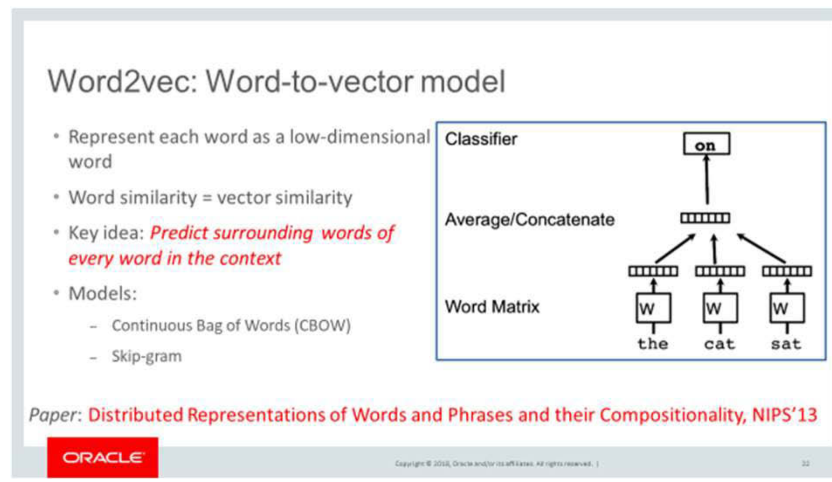
- There are several approaches now
  - (Academia and Industry)
  - An early approach that exploits techniques from modern NLP (natural language processing)
  - Word2Vec : a ML technique that learns closeness between words from large number of sentences
  - Perform many random walks on the graph
  - Apply W2V technique on random walk traces, treating vertices as words.

KDD'14


DeepWalk: Online Learning of Social Representations

Bryan Perozzi, Stony Brook University, Department of Computer Science
Rami Al-Rfou, Stony Brook University, Department of Computer Science
Steven Skiena, Stony Brook University, Department of Computer Science

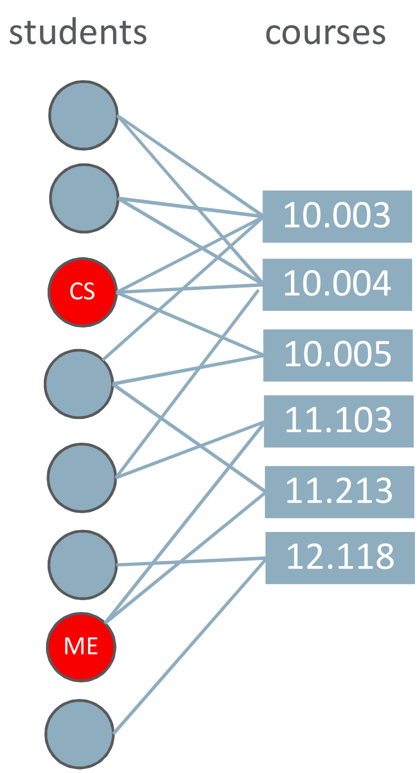
{bperozzi, ralrfou, skiena}@cs.stonybrook.edu



Example

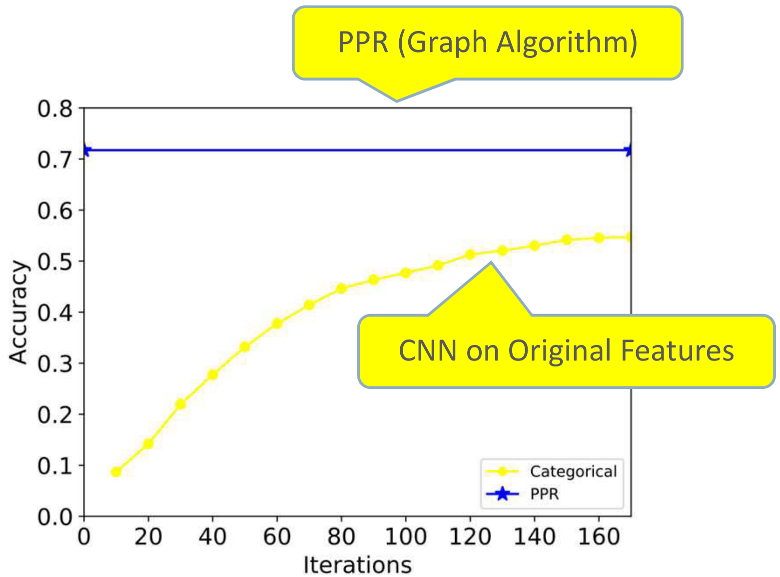
- Student classification
  - A dataset from university
  - Can you predict a student’s major or department just by looking at the classmates in the course that (s)he is taking?

- Note: you can consider this as an emulation of customer segmentation problem
  - Student => Customer
  - Course taking => Item or service purchase
  - Department => Segment label

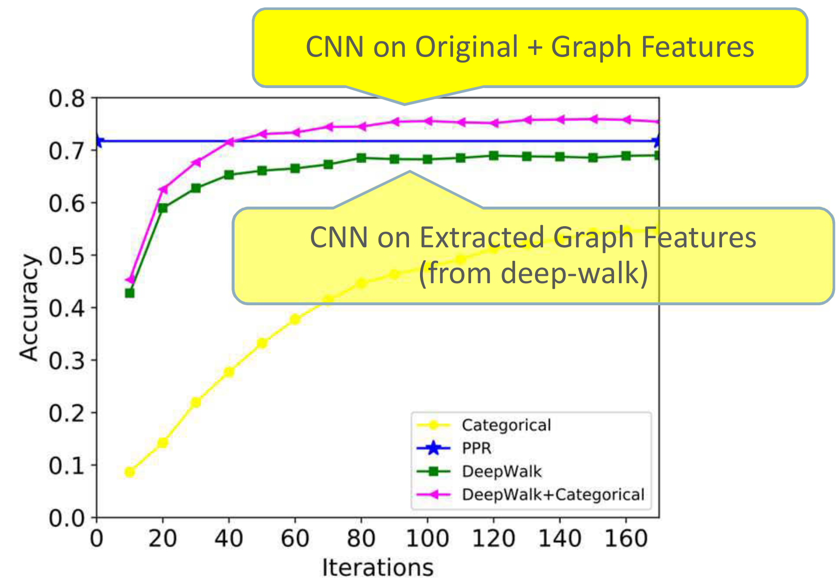


Results

- (Result #1) Graph-based prediction gives better result than naïve application of ML (e.g. CNN) on basic student features (e.g. age, gender, background, …)



- (Result #2) Deep-Walk preserves information from graph representation
- (Result #3) Deep-Walk allows to combined graph data with other features



ORACLE

Sounds complicated, how can I use this technique easily?

- We have an implementation in our graph package (PGX)
  - Load graph model
  - Compute graph embedding
  - Query embedding directly on graph
  - Export graph embedding

Shell Java

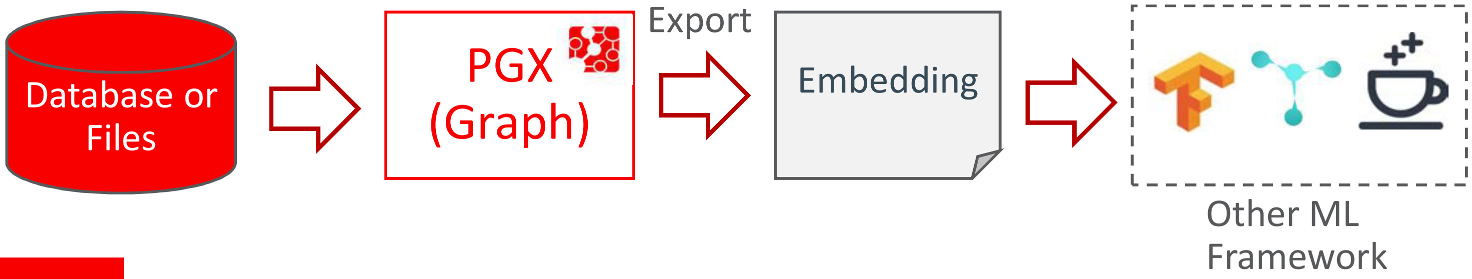
```
pgx> model = analyst.deepWalkModelBuilder().
        setMinWordFrequency(1).
        setBatchSize(512).
        setNumEpochs(1).
        setLayerSize(100).
        setLearningRate(0.05).
        setMinLearningRate(0.0001).
        setWindowSize(3).
        setWalksPerVertex(6).
        setWalkLength(4).
        setSampleRate(0.00001).
        setNegativeSample(2).
        setValidationFraction(0.01).
        build()
```

Training the DeepWalk model

We can train a `DeepWalk` model with the specified (defau

Shell Java

```
pgx> model.fit(graph)
```



ORACLE

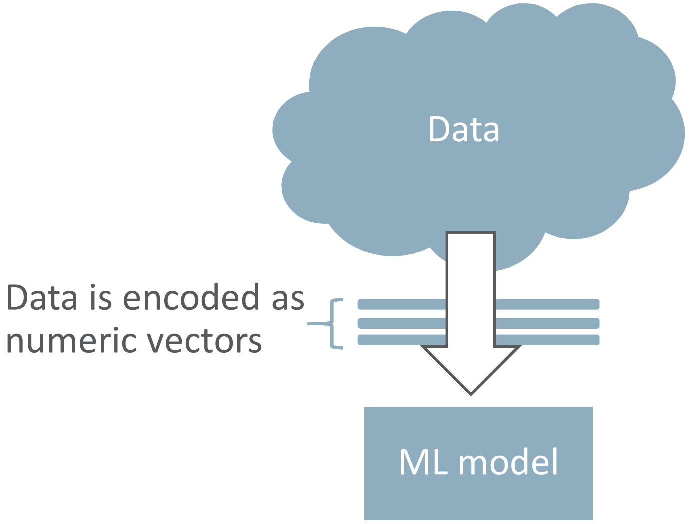
Agenda

1. Machine Learning and Graph Analysis
2. Encoding Relationship Distance
3. Encoding Irregular Structure (+ Demo)
4. Current Works and Directions

ORACLE

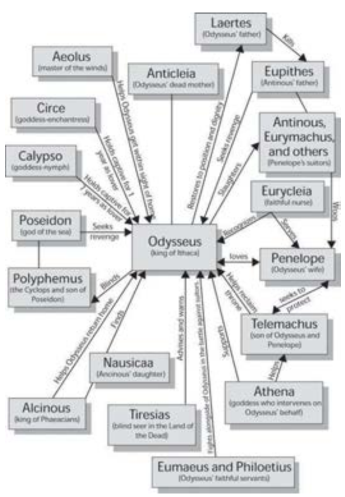
Yet Another Encoding Problem

- Again you want to
  - Capture relationships between entities and
  - Feed it into ML model
- But this time
  - Your focus is not an individual entity
  - Rather, you want to characterize *group of entities* that are related one another
  - Where these relationship structures are irregular and arbitrary
  - Still you want analyze these groups with ML
  - Challenge: how ‘irregular structures in entity relationships’ can be encoded for ML tasks?

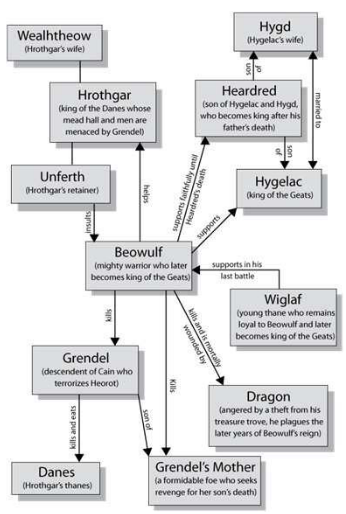


Classic Literature Example

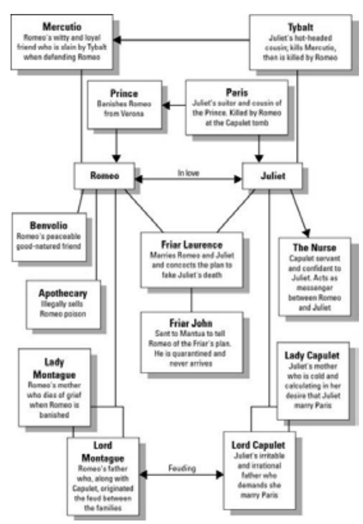
- Main character relationships in classic literature
  - Can we use ML to tell which pieces have similar character relationships? Cluster pieces by their similarity?
  - Given a modern piece, identify what classic piece has the most similar structure?



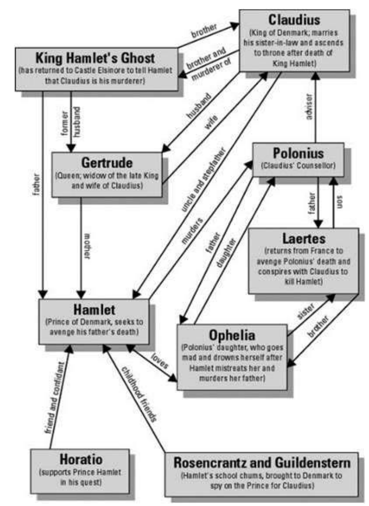
Odyssey



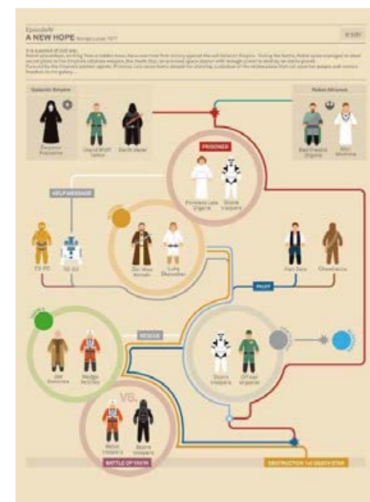
Beowulf



Romeo & Juliet



Hamlet



Star Wars: Episode IV

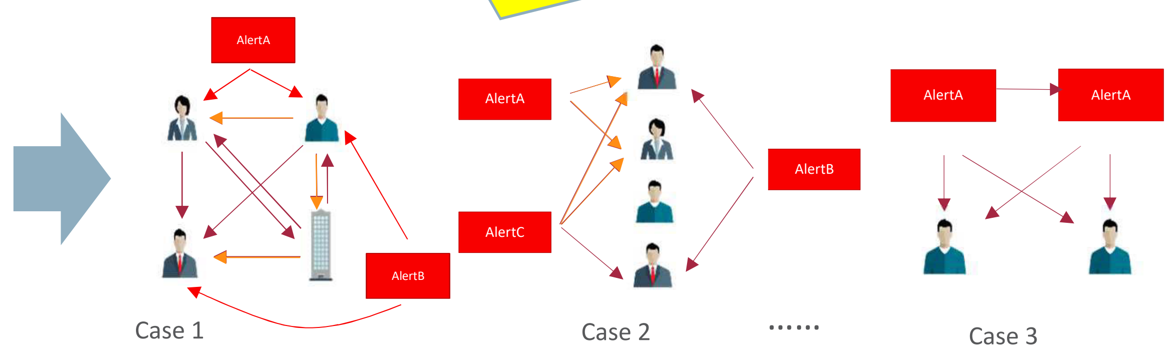
More Serious Example – Anti-Money Laundering

- An application from financial domain
  - Real-time monitoring of financial transactions
  - ➔ Any suspicious transactions are *tagged* as an event (but with a lot of false-positives)
  - Correlated events are gathered up to form a case
  - ➔ Each case is put under investigation (by experts)
  - ➔ A case can end up with either real or false positive
-

Can we train a ML model to distinguish real money laundering from false positives?
How to encode these structures?

Originator	Beneficiary	Time Stamp	Amount	...
Paul	Zion Bank	6/19 5:59:59.336 UTC+8	$ 30,000	...
Jack	E-Weddingbands LLC	6/19 5:59:59.516 UTC+8	$ 112,000	...
James	Provo Bank	6/19 6:01:20.222 UTC+8	$ 150.00	...
Steve	Linda	6/19 6:02:55.222 UTC+8	$ 999.30	...
...	...	...	...	...

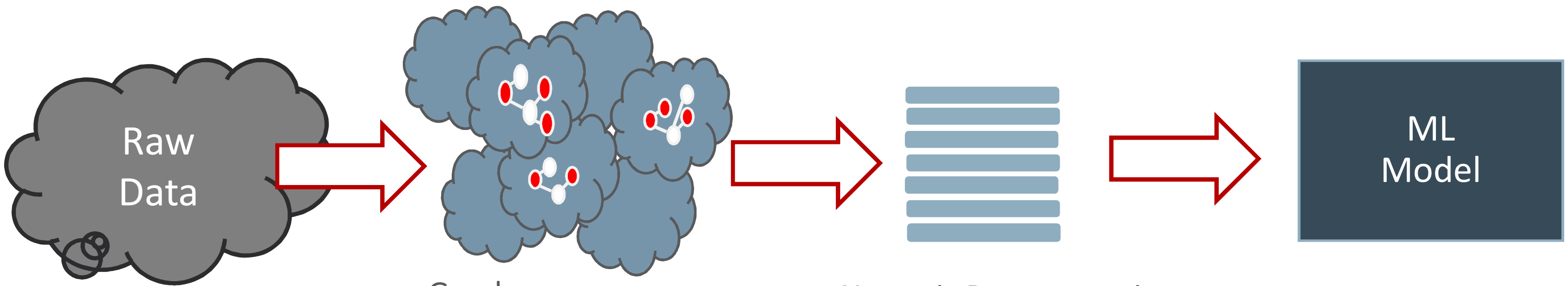
Real-Time Transaction Monitoring



ORACLE

Problem Definition Again

- Now the dataset is a set of (relatively small) graphs
- We would like to find N-dimensional vector representation of each graph such that
- If two graph G1, G2 are *similar in shape*, ‖ *v(G1) - v(G2)* ‖ small in the N-dimensional vector space

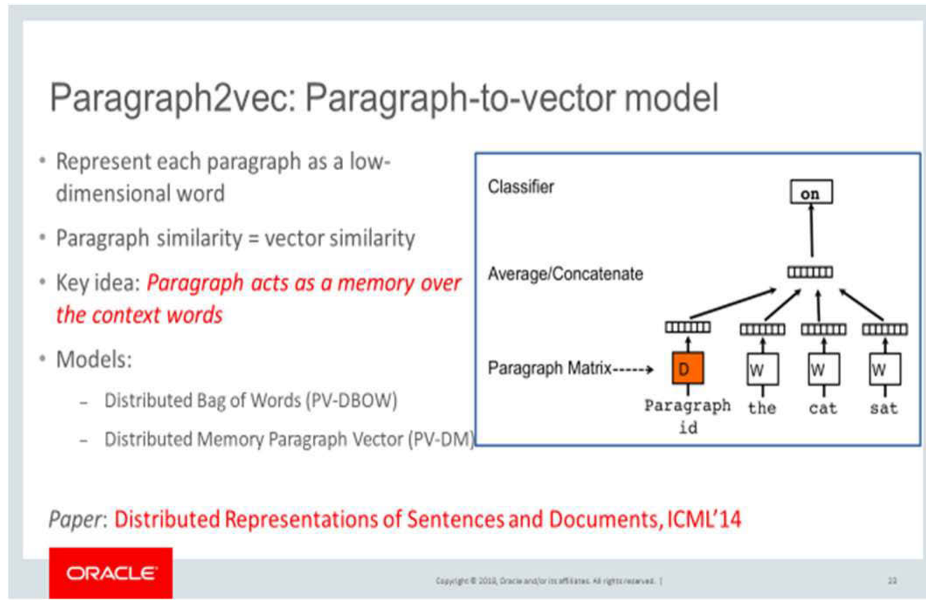


Graph Representations

Numeric Representation (N-dimensional vector)

Approach

- Again, adopt a technique from NLP, or sequence-based learning
  – i.e: Paragraph2Vec ➔ learn from large text corpus. Paragraphs composed of *similar* words are close in embedding space
  – Consider each graph as paragraph
  – Generate random-walk on each graphs
  – Apply Paragraph2Vec model (each graph become a paragraph)



Adding Secret Sauces

- We applied some of our own techniques
  - better quality of answer than naïve application of paragraph2vec
  - (1) Consider multiple properties (rather than single label)
  - (2) When generating traces, consider edges (instead of vertices) as words.
  - (3) Attach global properties of the each graph -- e.g. size of graph

Sounds complicated, how can I use this technique easily? (2)

- We have an implementation in our graph package (PGX)
  - Load data – unconnected graphs
  - Compute graph embedding
  - Query embedding directly or
  - Export graph embedding

Shell | Java

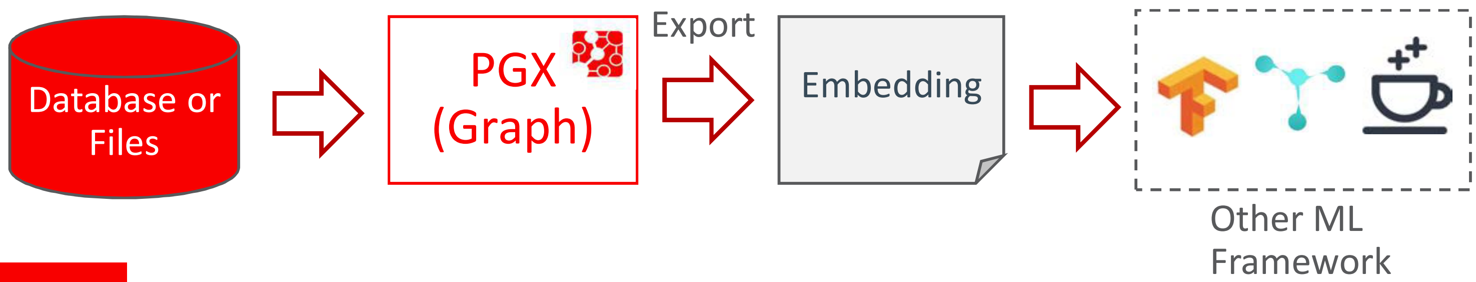
```
pgx> model = analyst.pg2vecModelBuilder().
         setGraphLetIdPropertyName("graph_id").
         setVertexPropertyNames(Arrays.asList("category")).
         setMinWordFrequency(1).
         setBatchSize(128).
         setNumEpochs(5).
         setLayerSize(200).
         setLearningRate(0.04).
         setMinLearningRate(0.0001).
         setWindowSize(4).
         setWalksPerVertex(5).
         setWalkLength(8).
         setUseGraphletSize(true).
         setValidationFraction(0.05).
         build()
```

Training the Pg2vec model

We can train a `Pg2vec` model with the specified (default or customized) settings a

Shell | Java

```
pgx> model.fit(graph)
```



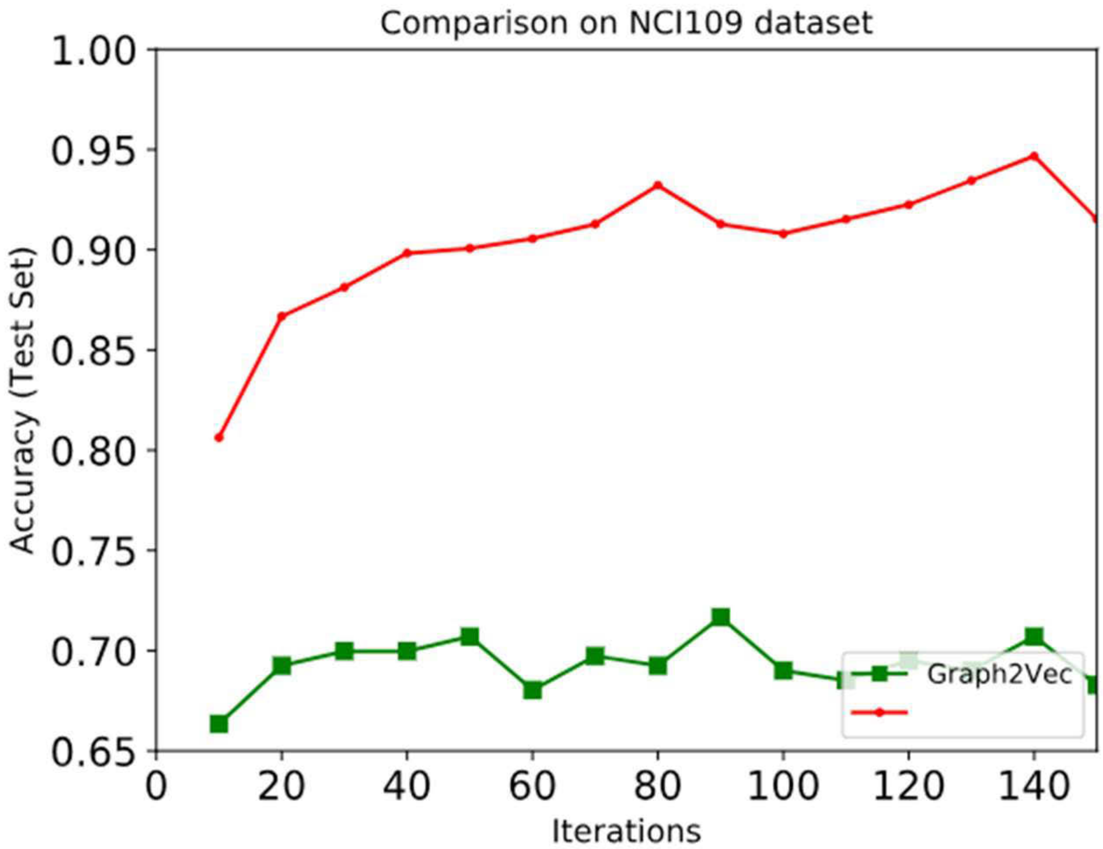
Evaluation

- **Datasets: cheminformatics**
  - National Cancer Institute (NCI109)
    - #Graphs: 4127
      - #Vertices: ranges from 35 to 111
      - #Edges: ranges from 152 to 476
    - *Cancer types (binary classification)*
  - Proteins
    - #Graphs: 1113
      - #Vertices: ranges from 9 to 620
      - #Edges: ranges from 64 to 4048
    - *Protein types (binary classification)*

By looking at the molecular structure as graph, classify whether it is cancerous or not

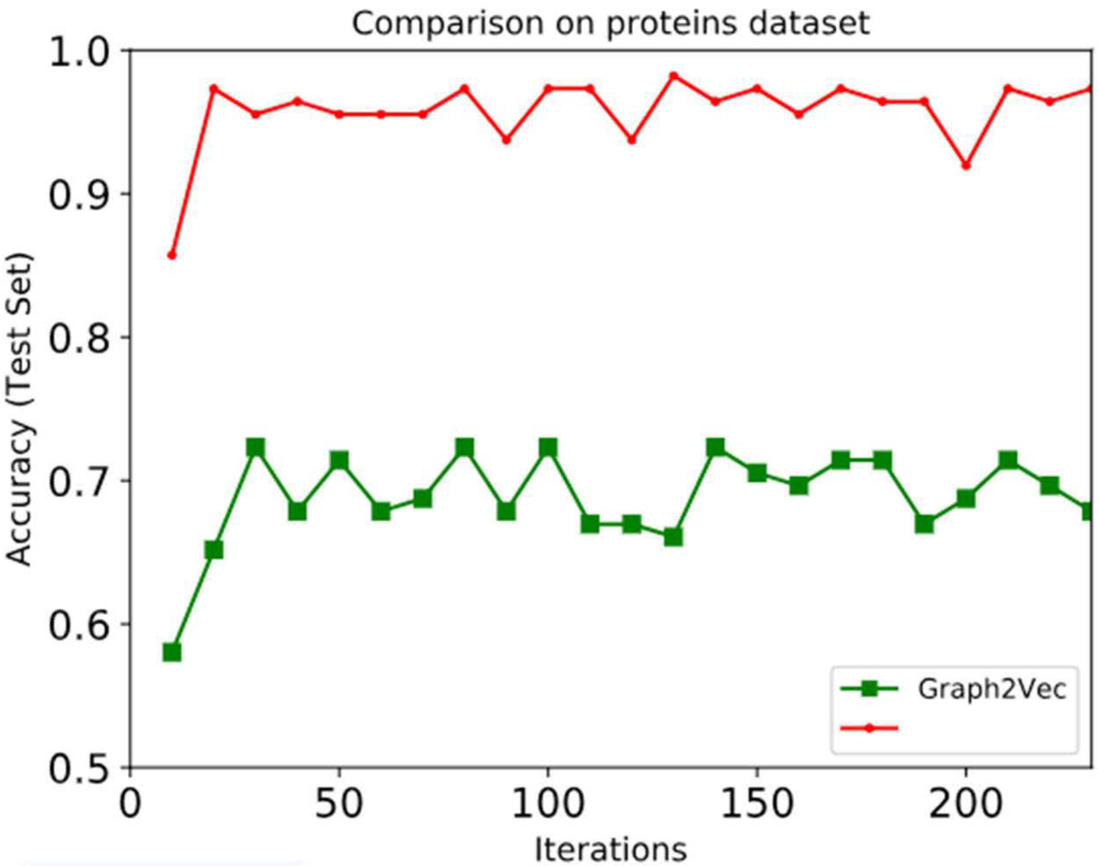
ORACLE

Classification task: NCI109 dataset



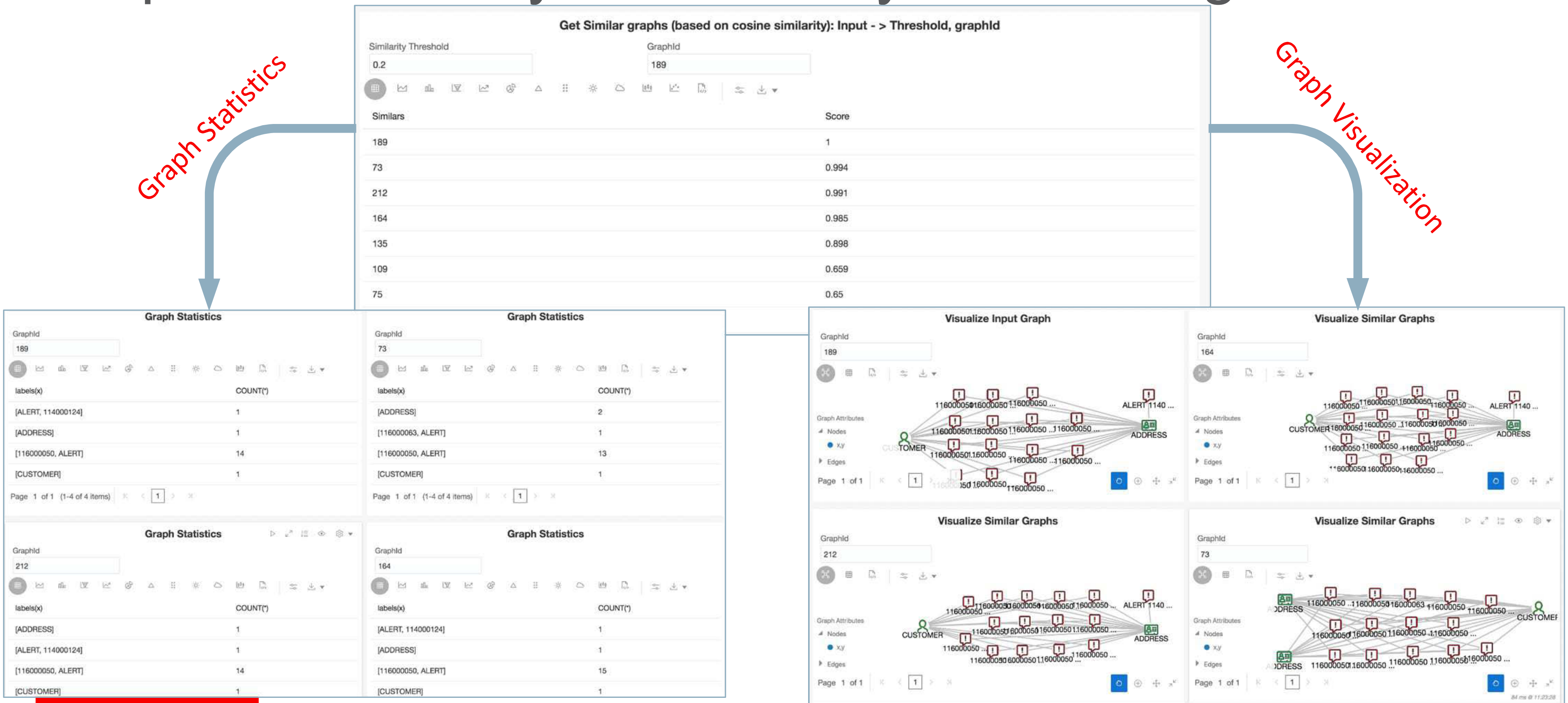
- Train:test = 9:1
- Quality improvement
  - ~22% in classification acc.
- Comparison:
  - Graph2vec (MLG'17)
  - Similar approach to ours without the secret sauce

Classification task: Proteins dataset



- Train:test = 9:1
- Quality improvement
  - ~25% in classification acc.

GraphLet similarity: Anti-Money Laundering dataset



Similars	Score
189	1
73	0.994
212	0.991
164	0.985
135	0.898
109	0.659
75	0.65

Graph Statistics (GraphId 189)

labels(x)	COUNT(*)
[ALERT, 114000124]	1
[ADDRESS]	1
[116000050, ALERT]	14
[CUSTOMER]	1

Graph Statistics (GraphId 73)

labels(x)	COUNT(*)
[ADDRESS]	2
[116000063, ALERT]	1
[116000050, ALERT]	13
[CUSTOMER]	1

Graph Statistics (GraphId 212)

labels(x)	COUNT(*)
[ADDRESS]	1
[ALERT, 114000124]	1
[116000050, ALERT]	14
[CUSTOMER]	1

Graph Statistics (GraphId 164)

labels(x)	COUNT(*)
[ALERT, 114000124]	1
[ADDRESS]	1
[116000050, ALERT]	15
[CUSTOMER]	1

 | Rhicheek Patra, Oracle Labs (ML Summit 2018)

Demo

Demo (Recording)

Agenda

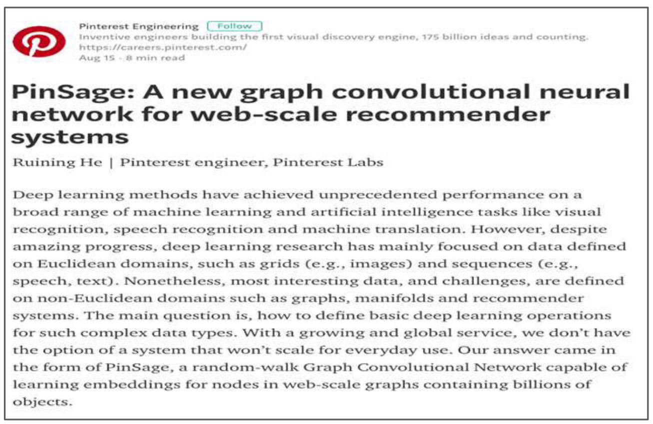
1. Machine Learning and Graph Analysis
2. Encoding Relationship Distance
3. Encoding Irregular Structure (+ Demo)
4. Current Works and Directions

ORACLE

Graphs and Machine Learning

- By the way, combining graph and machine learning is a trend
  - Many in industry and academia are looking at this problem
  - And applying it to solving real problems

Pintrest

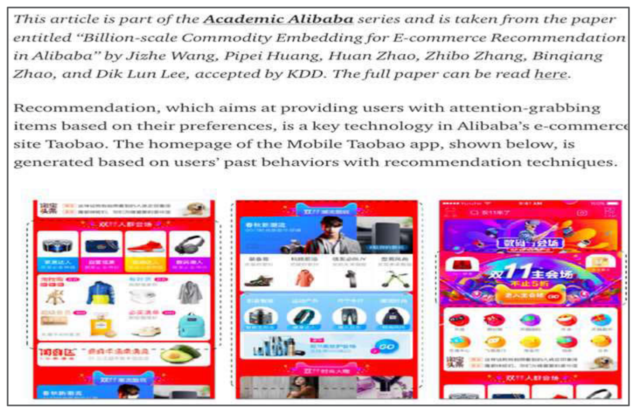

Pinterest Engineering Follow
Inventive engineers building the first visual discovery engine, 175 billion ideas and counting.
https://careers.pinterest.com/
Aug 15 · 8 min read

PinSage: A new graph convolutional neural network for web-scale recommender systems

Ruining He | Pinterest engineer, Pinterest Labs

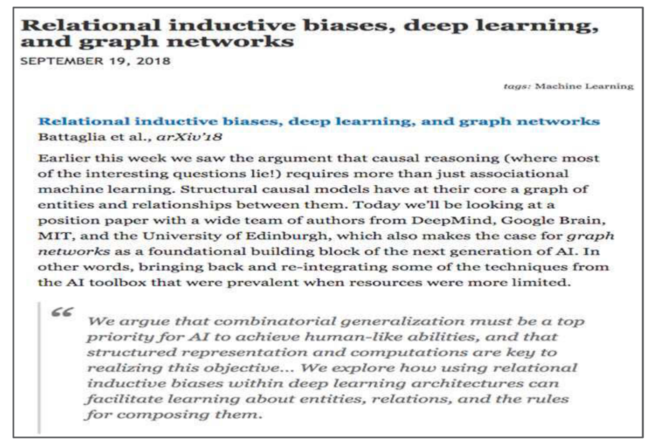
Deep learning methods have achieved unprecedented performance on a broad range of machine learning and artificial intelligence tasks like visual recognition, speech recognition and machine translation. However, despite amazing progress, deep learning research has mainly focused on data defined on Euclidean domains, such as grids (e.g., images) and sequences (e.g., speech, text). Nonetheless, most interesting data, and challenges, are defined on non-Euclidean domains such as graphs, manifolds and recommender systems. The main question is, how to define basic deep learning operations for such complex data types. With a growing and global service, we don't have the option of a system that won't scale for everyday use. Our answer came in the form of PinSage, a random-walk Graph Convolutional Network capable of learning embeddings for nodes in web-scale graphs containing billions of objects.

Alibaba


*This article is part of the **Academic Alibaba** series and is taken from the paper entitled "Billion-scale Commodity Embedding for E-commerce Recommendation in Alibaba" by Jizhe Wang, Pipei Huang, Huan Zhao, Zhibo Zhang, Binqiang Zhao, and Dik Lun Lee, accepted by KDD. The full paper can be read here.*

Recommendation, which aims at providing users with attention-grabbing items based on their preferences, is a key technology in Alibaba's e-commerce site Taobao. The homepage of the Mobile Taobao app, shown below, is generated based on users' past behaviors with recommendation techniques.

Google


Relational inductive biases, deep learning, and graph networks

SEPTEMBER 19, 2018

tags: Machine Learning

Relational inductive biases, deep learning, and graph networks
Battaglia et al., *arXiv'18*

Earlier this week we saw the argument that causal reasoning (where most of the interesting questions lie!) requires more than just associational machine learning. Structural causal models have at their core a graph of entities and relationships between them. Today we'll be looking at a position paper with a wide team of authors from DeepMind, Google Brain, MIT, and the University of Edinburgh, which also makes the case for *graph networks* as a foundational building block of the next generation of AI. In other words, bringing back and re-integrating some of the techniques from the AI toolbox that were prevalent when resources were more limited.

> *We argue that combinatorial generalization must be a top priority for AI to achieve human-like abilities, and that structured representation and computations are key to realizing this objective... We explore how using relational inductive biases within deep learning architectures can facilitate learning about entities, relations, and the rules for composing them.*

Other use cases

- In general, graph analysis can be combined with ML for various applications
- Where analyzing entity-entity relationships is required
  - Personalized Recommendation, Customer segmentation, …
  - Fraud Detection – Health care, Insurance, …
  - Cyber Security – Network Intrusion
  - SNS analysis
  - Fake new detections
  - …

ORACLE

Directions

- Improving Scalability
  - Increasing the size of graph (e.g. tens of billions of vertices)

- Combining structure (relationship) and other raw observation
  - E.g. Item attributes + Co-purchase Information
  - Finding more elegant solution than simple ensemble techniques

ORACLE

Summary

- Modern techniques for combining Graph Analysis and Machine Learning
- Graph captures fine-grained relationships between data entities
- Adopt techniques from NLP to encode relationship information
  - Vertex2Vector (capture relationship-induced distance between entities)
  - Graph2Vector (capture similarities between graph instances)
- Applicable to many real-world applications
- Implementation (soon) available in Oracle's graph package

ORACLE

Related Talks

Date and Time	Location	Title
Monday, 9:00 a.m. - 9:45 a.m.	Moscone West - **Room 2016**	Graph Query Language For Navigating Complex Data [DEV5447]
Monday, 10:30 a.m. - 11:15 a.m	Moscone West - **Room 2022**	When Graphs Meet Machine Learning [DEV5420]
Monday, 11:30 a.m. - 12:15 p.m.	Moscone West - **Room 2022**	Automate Anomaly Detection with Graph Analytics [DEV5397]
Monday, 12:30 p.m. - 1:15 p.m.	Moscone West - **Room 2022**	Oracle Database MLE: JavaScript, Python, and More *in* the Database [DEV5082]
Monday, 1:30 p.m. - 2:15 p.m..	Moscone West - **Room 2003**	How to Build Geospatial Analytics with Python and Oracle Database [DEV5185]
Monday, 4:45 p.m. - 5:30 p.m.	Moscone West - **Room 3004**	Introduction to Graph Analytics and Oracle Graph Cloud Service [TRN4098]
Monday, 5:45 p.m. - 6:30 p.m.	Moscone West - **Room 3004**	How to Analyze Data Warehouse Data as a Graph [TRN4099]
Thursday, 2:00 p.m. - 2:45 p.m.	Moscone West - **Room 2018**	Analyzing Blockchain and Bitcoin Transaction Data as Graphs [DEV4835]

Date and Time	Location	Title (Meet The Experts)
Tuesday, noon - 1:00 p.m.	Moscone West – **Lounge B**	Graph Analysis and Database Technologies
Tuesday, 3:00 p.m. – 4:00 p.m	Moscone West – **Lounge B**	Graph Analysis and Machine Learning (Graph Queries and Analysis)
Wednesday, 10:00 a.m. - 11:00 a.m.	Moscone West - **Lounge B**	Graph Analysis and Database Technologies
Wednesday, 11:00 a.m. - noon	Moscone West - **Lounge A**	Graph Analysis and Machine Learning (Graph Queries and Analysis)

ORACLE

Spatial and Graph at OOW 2018

View this list at tinyurl.com/SpatialGraphOOW18

Demos

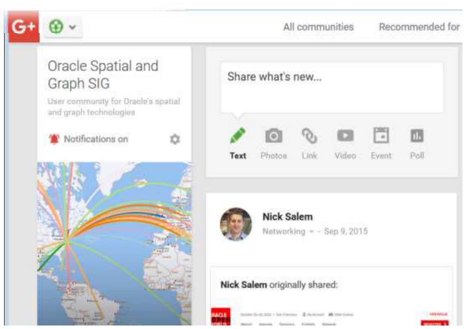
Date/Time	Title	Location
• Monday 9:45 am – 5:45 pm • Tuesday 10:30 am – 5:45 pm • Wednesday 10:30 am– 4:45 pm	Oracle Spatial and Graph Database, Analytics, and Cloud Services  	Moscone South Exhibit Hall (‘The Exchange’) • Oracle Demogrounds: Cloud Platform > Application Development area • Kiosk # APD-WU3

ORACLE

The Spatial & Graph SIG User Community

We are a vibrant community of customers and partners that connects and exchanges knowledge online, and at conferences and events.

Oracle Spatial & Graph SIG



Join us online

tinyurl.com/oraclespatialcommunity

LinkedIn Google+ IOUG IOUG SIG

@oraspatialsig

oraclespatialsig@gmail.com

ORACLE

Call for Speakers now open!
Submit an abstract to share your use case or technical session

Analytics and Data Summit

All Analytics. All Data. No Nonsense.

March 12 – 14, 2019

BIWA

Formerly called the BIWA Summit with the Spatial and Graph Summit
Same great technical content…new name!
www.AnalyticsandDataSummit.org

Oracle Spatial & Graph SIG

ORACLE®