

VIGOR: Interactive Visual Exploration of Graph Query Results (VAST 2017)

ROBERT PIENTA, FRED HOHMAN, ALEX ENDERT, ACAR TAMERSOY, KEVIN ROUNDY, CHRIS GATES,
SHAMKANT NAVATHE, DUEN HORNG CHAU

PRESENTED BY: RITESH SARKHEL, OMID ASUDEH, MONIBA KEYMANESH

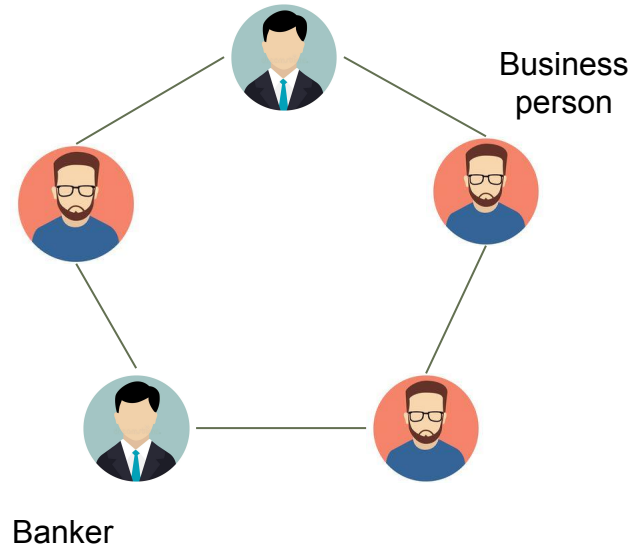


Outline

- ❑ Background
- ❑ Motivation
- ❑ VIGOR Interface Overview
- ❑ Design Rationale
- ❑ Methodology
- ❑ Experimental Results
- ❑ Pros/Cons
- ❑ Demo of the tool

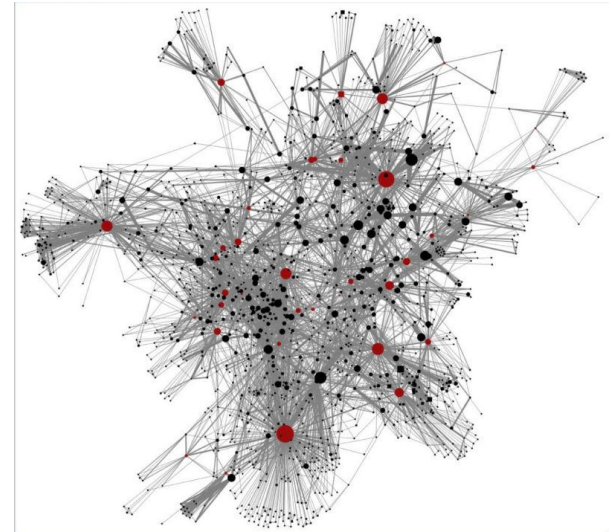
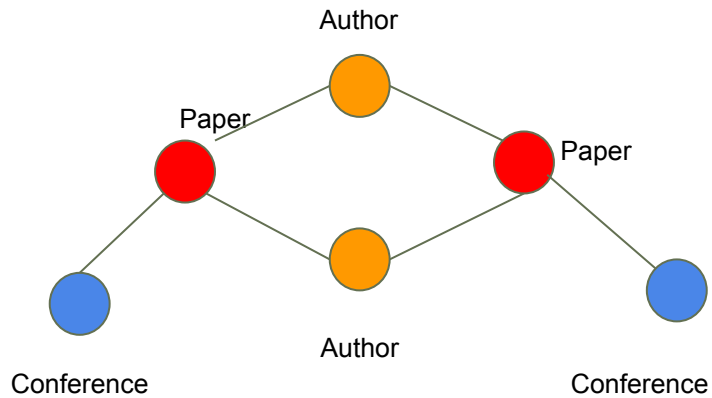
Background

- ❏ What is graph querying?
- ❏ Why do we query graphs?
 - ❏ Detecting money laundering rings
 - ❏ Discovering near bipartite cores in auction fraud
 - ❏ Uncovering fraudulent near-clique reviews



Background

Analyst



Results: subgraph matches

Motivation

- ❑ Different node features
- ❑ Arbitrary user queries
- ❑ Large number of results
- ❑ Shared nodes and edges among the results

Cannot not help understand underlying patterns



Explainability

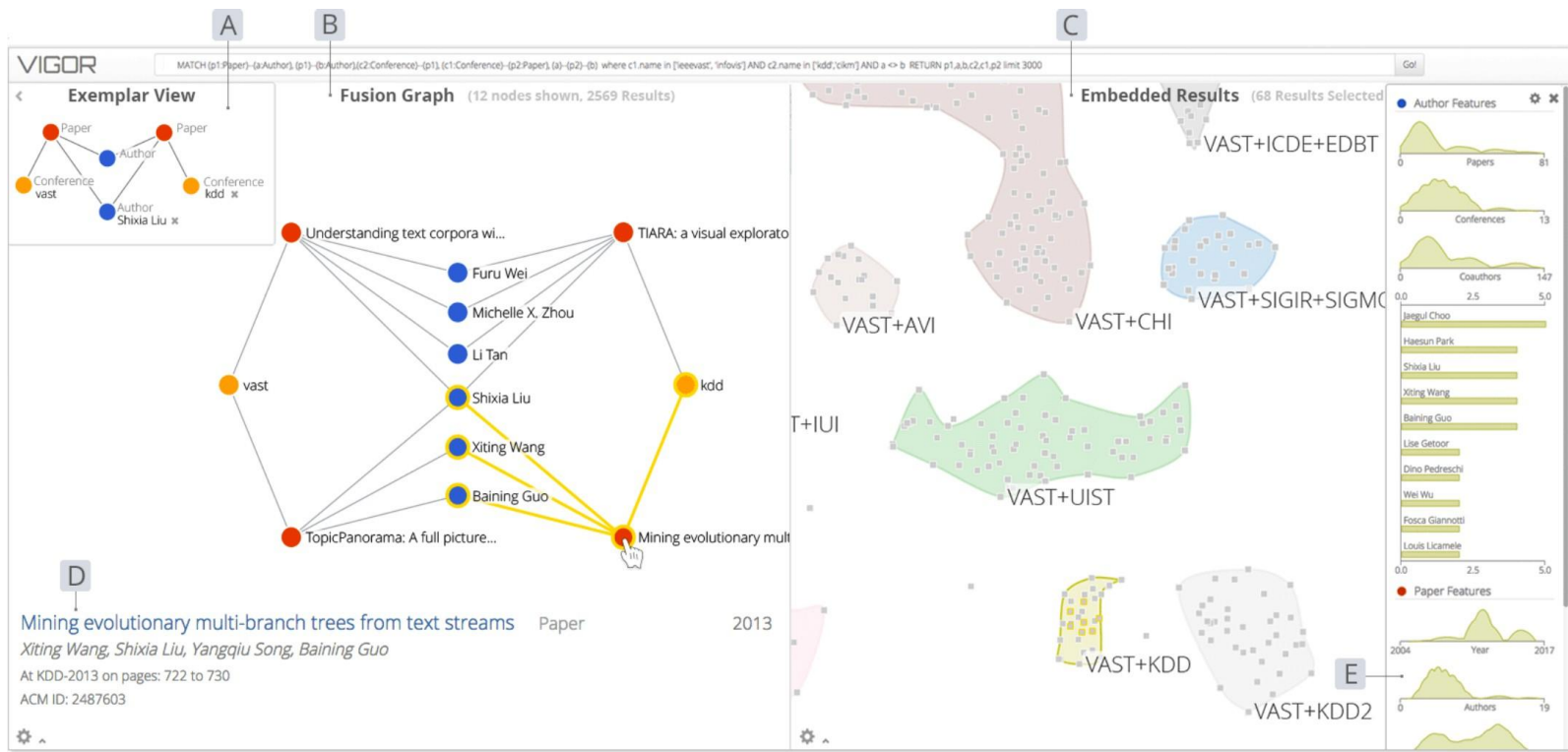
Cannot explore the result set



Interactive query expansion/filtering

Solution: **VIGOR** Graph Query Visualization and Exploration System

Overview of VIGOR



Interface Overview

VIGOR

Exemplar view

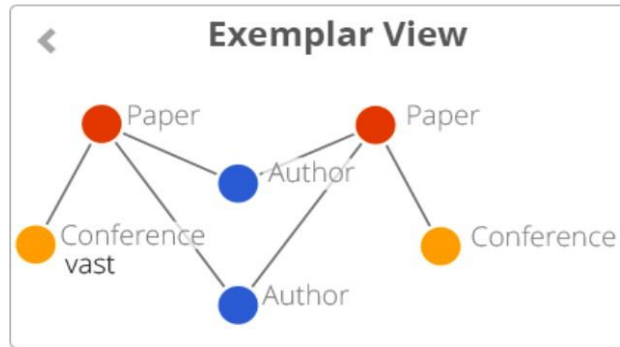
Fusion graph

Subgraph
Embedding View

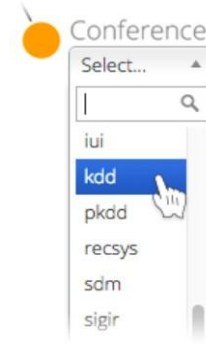
Feature Explorer
View

Exemplar view

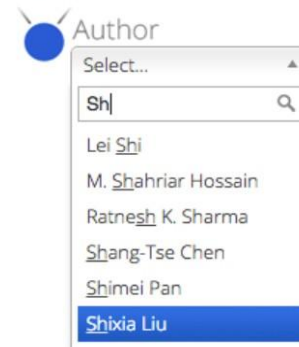
A Visualized Query



B Selecting KDD



C Selecting Shixia Liu



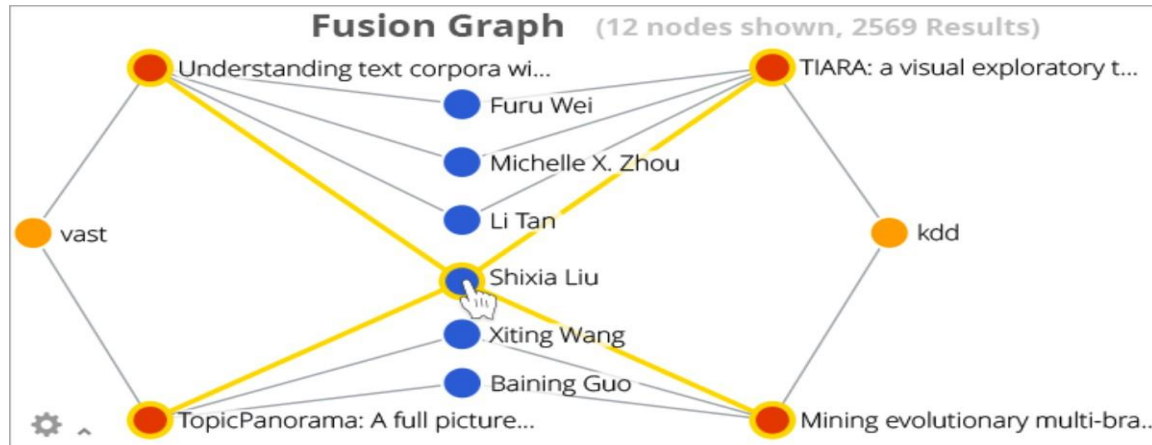
Alice enters a graph query



Cypher QL transforms query to a graph

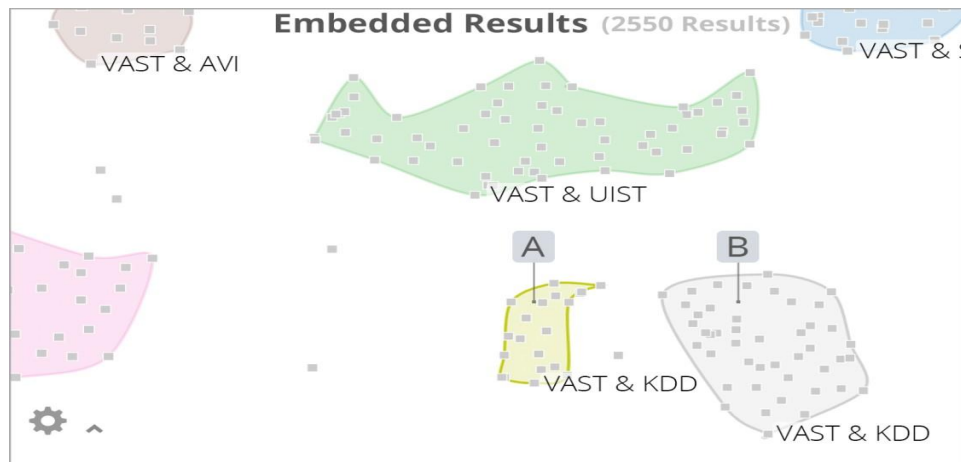
"Find authors who have papers both in VAST and KDD"

Fusion graph



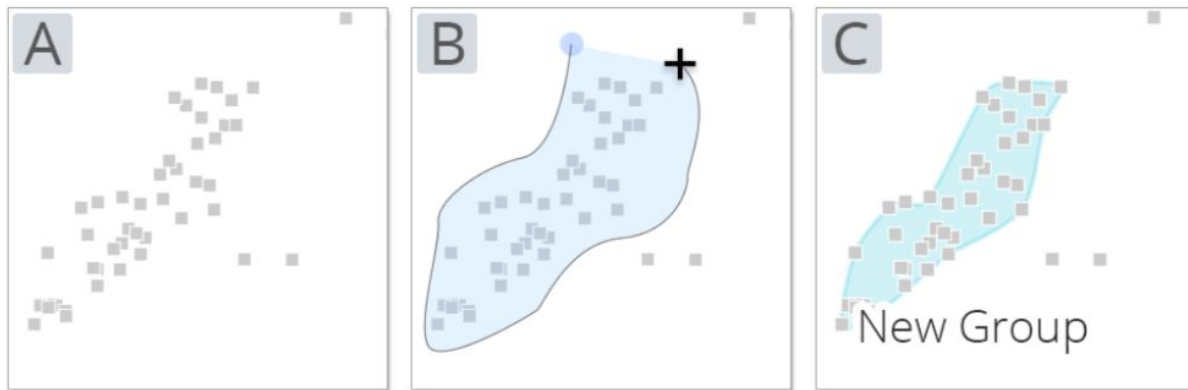
- VIGOR provides a hybrid view of how the query results match up with the input query
- Selecting a result node highlights the relevant attributes related to that tuple

Subgraph Embedding View



- Each square point represents a node in result-subgraph
- Similar results are spatially close in embedding space
- Points in same cluster are bounded by a concave hull

Subgraph Embedding View



Alice's graph query
on DBLP



~ 2500 results of
(paper, author-list)

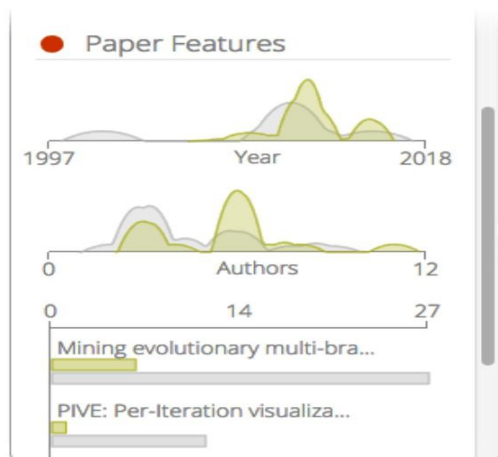
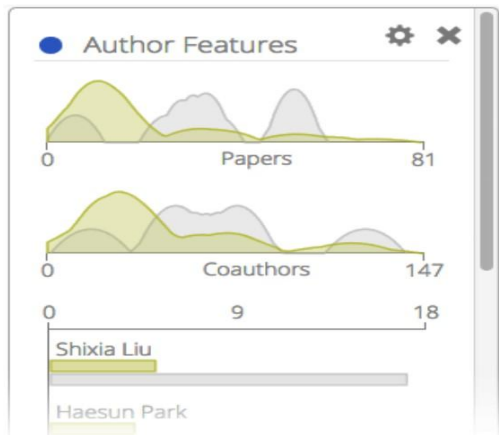


Result subgraphs
clustered in a
feature aware space



Considers all possible
combinations of
features and PCA to a
two-dimensional space

Feature Explorer View



- VIGOR provides the feature-level flexibility
- Subgraph Embedding changes based on modified features
- Supports both continuous and discrete feature values

Design Rationale

□ Exemplar View provides:

- ✓ Fast error-checking for input query
- ✓ Bottom-up exploration of result-set
- ✓ Ability to start from a familiar result and on the fly query expansion or filtering

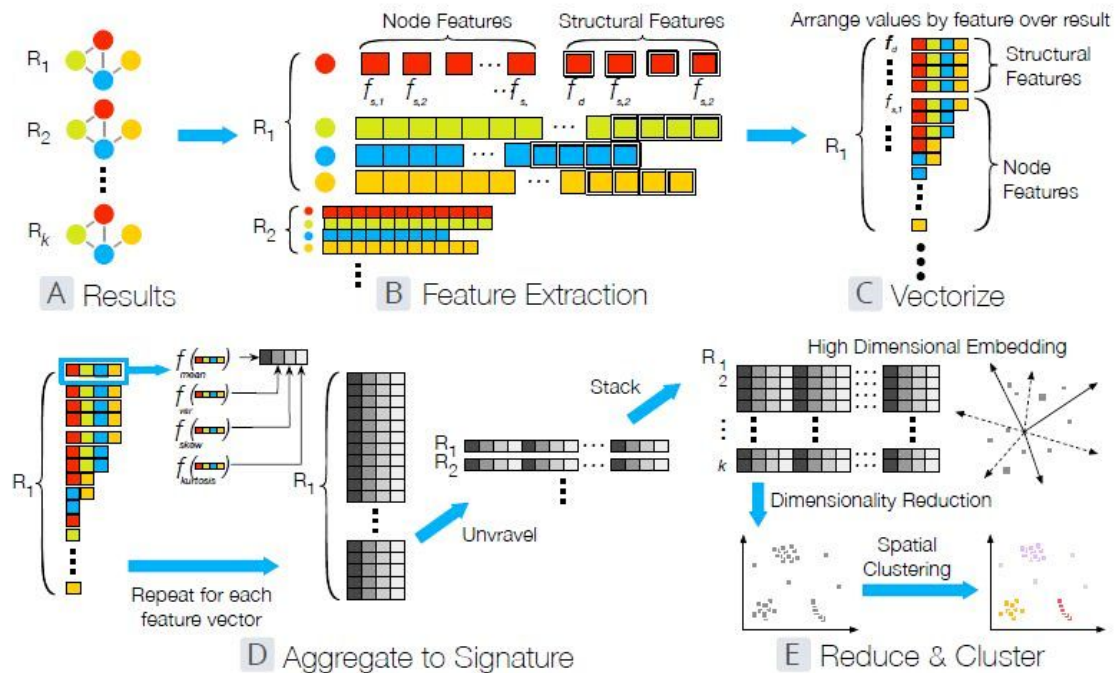
□ Subgraph Embedding provides:

- ✓ Top-down exploration of the result-set
- ✓ Global view of the result set
- ✓ Clustering similar results together help analysts get a macro-view

Design Rationale

- ❑ Feature-centric sense-making by Feature Explorer View
- ❑ VIGOR provides coordination among multiple views
 - ✓ Fusion graph combines top-down and bottom-up views of result set
 - ✓ Clicking on the node explains why that result was included
 - ✓ Hovering over the squares on Subgraph Embedding View shows detailed values

Methodology



Methodology contd.

1. Extract - Structural/Node features for each node

✓ Structural features:

- Node degree
- Egonet edges
- Egonet neighboring nodes
- Clustering coefficients

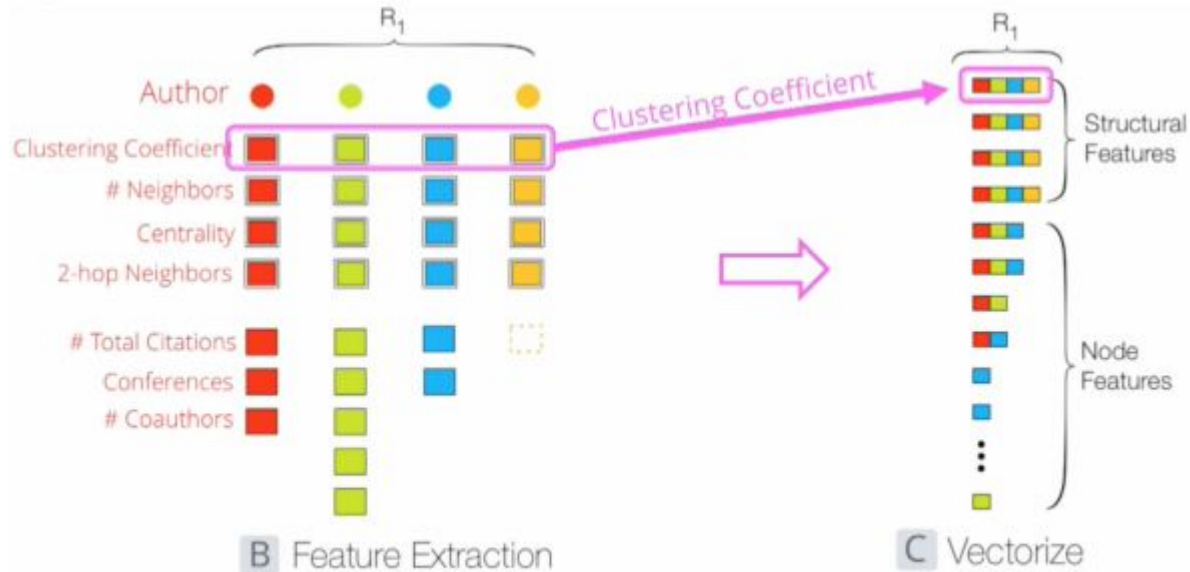
✓ Node features:

- For author: number of co-authors, number of conferences

2. Vectorize - Merge the common features into per-result vectors

3. Aggregate & Normalize into Signature- by computing mean, variance, skewness and kurtosis for each feature for each node .

Merge common feature



Methodology contd.

4.Reduction and Clustering

- ✓ Reduce dimensionality to 2 using PCA / kernel-PCA / t-SNE
- ✓ Cluster results based on feature combinations using OPTICS
 - Canberra distance (weighted version of manhattan distance)

$$d(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^n \frac{|p_i - q_i|}{|p_i| + |q_i|}$$

where

$$\mathbf{p} = (p_1, p_2, \dots, p_n) \text{ and } \mathbf{q} = (q_1, q_2, \dots, q_n)$$

- Canberra distance is sensitive to small changes near zero, which helps preserves small distances in the final reduction.

Architecture

- ❑ Client-server architecture using D3 and jQuery
- ❑ Backend in python

Experimental results

- User study
- Think-aloud explorative study

User Study 1: DBLP

Within-subject User Study

Participants: 12 (7 female, 5 male; ages 21 to 31)

Dataset: DBLP Co-authorship Network

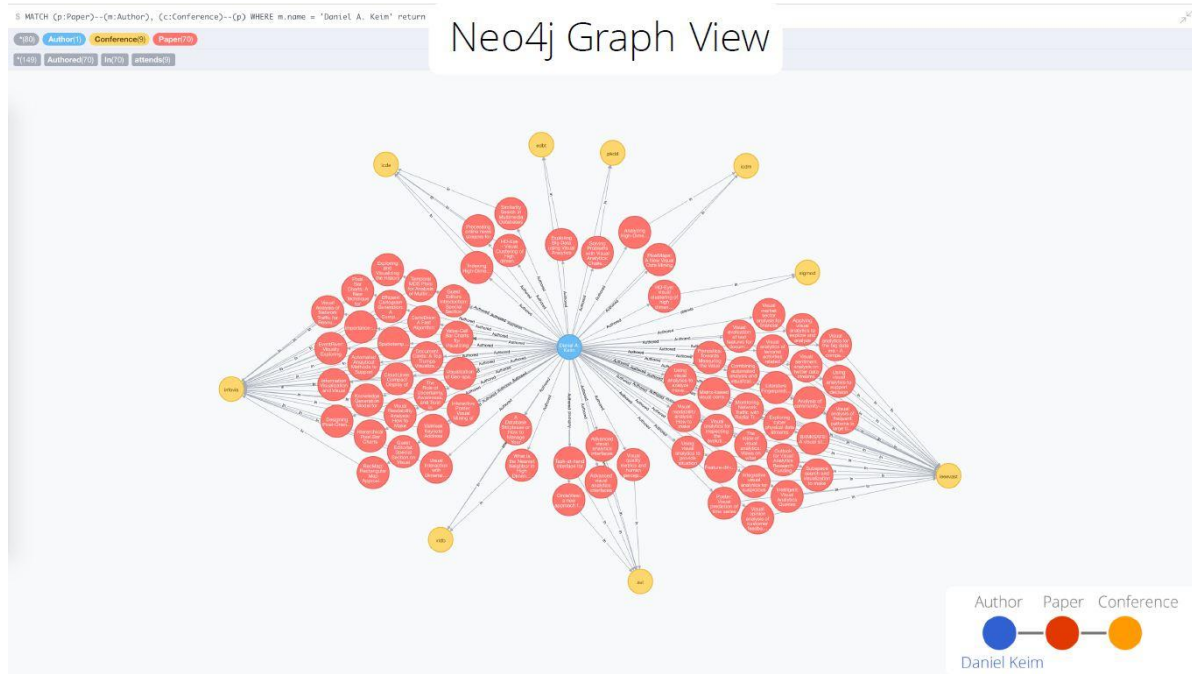
Tasks: 4 tasks related to co-authorship and conferences

Measured: Task completion times and error rates



Slides adapted from Robert Pienta et al. @GaTech with the authors' permission

Experimental results (contd.)



Slides adapted from Robert Pienta et al. @GaTech with the authors' permission

Experimental results (contd.)

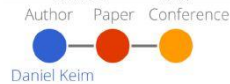
```

MATCH (p:Paper)--(a:Author), (c:Conference)--(p) WHERE a.name = 'Daniel A. Keim' return
    
```

Paper Title
Outlook for Visual Analytics Research Funding
Using visual analytics to support decision making to solve the Kronos incident (VAST challenge 2014)
A Database Striptease or How to Manage Your Personal Databases
Guest Editors Introduction: Special Section on InfoVis
Visual opinion analysis of customer feedback data
Analysis of community-contributed space- and time-referenced data (example of flickr and panorama photos)
Exploring and Visualizing the History of InfoVis
Visual quality metrics and human perception: an initial study on 2D projections of large multidimensional data
Using visual analytics to provide situation awareness for movement and communication data
CircleView: a new approach for visualizing time-related multidimensional data sets
Applying visual analytics to explore and analyze movement data
Poster: Visual prediction of time series
Visual sentiment analysis on twitter data streams
What is the Nearest Neighbor in High Dimensional Spaces
Visualization of Geo-spatial Point Sets via Global Shape Transformation and Local Pixel Placement
Monitoring Network Traffic with Radial Traffic Analyzer
Integrative visual analytics for suspicious behavior detection
CloudLines: Compact Display of Event Episodes in Multiple Time-Series
Pinotics: Towards Measuring the Value of Visualization
Temporal MDS Plots for Analysis of Multivariate Data
Similarity Search in Multimedia Databases
The Role of Uncertainty, Awareness, and Trust in Visual Analytics
Intelligent Visual Analytics Queries
Processing online news streams for large-scale semantic analysis
Literature Fingerprinting: A New Method for Visual Literary Analysis
Analyzing High-Dimensional Data by Subspace Validity
Task-at-hand interface for change detection in stock market data
VisWeek Keynote Address
Interactive Poster: Visual Mining of Business Process Data
PixelMaps: A New Visual Data Mining Approach for Analyzing Large Spatial Data Sets
Visual Interaction with Dimensionality Reduction: A Structured Literature Analysis
Combining automated analysis and visualization techniques for effective exploration of high-dimensional data
Guest Editorial: Special Section on Visual Analytics
Value-Cell Bar Charts for Visualizing Large Transaction Data Sets
Exploring Big Data using Visual Analytics
The state of visual analytics: Views on what visual analytics is and where it is going

Neo4j Table View

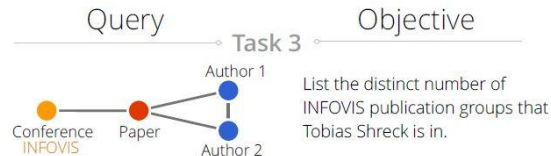
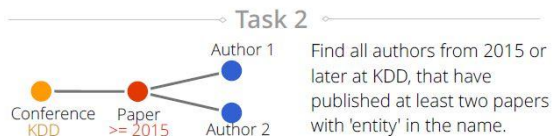
Name	Conference
Daniel A. Keim	isee'vat
Daniel A. Keim	isee'vat
Daniel A. Keim	vidb
Daniel A. Keim	infovis
Daniel A. Keim	isee'vat
Daniel A. Keim	isee'vat
Daniel A. Keim	infovis
Daniel A. Keim	avi
Daniel A. Keim	isee'vat
Daniel A. Keim	avi
Daniel A. Keim	isee'vat
Daniel A. Keim	isee'vat
Daniel A. Keim	isee'vat
Daniel A. Keim	vidb
Daniel A. Keim	infovis
Daniel A. Keim	isee'vat
Daniel A. Keim	isee'vat
Daniel A. Keim	isee'vat
Daniel A. Keim	infovis
Daniel A. Keim	isee'vat
Daniel A. Keim	icode
Daniel A. Keim	infovis
Daniel A. Keim	isee'vat
Daniel A. Keim	icode
Daniel A. Keim	isee'vat
Daniel A. Keim	icdm
Daniel A. Keim	avi
Daniel A. Keim	infovis
Daniel A. Keim	infovis
Daniel A. Keim	icdm
Daniel A. Keim	infovis
Daniel A. Keim	isee'vat



Slides adapted from Robert Pienta et al. @GaTech with the authors' permission

Experimental results (contd.)

User Study Tasks

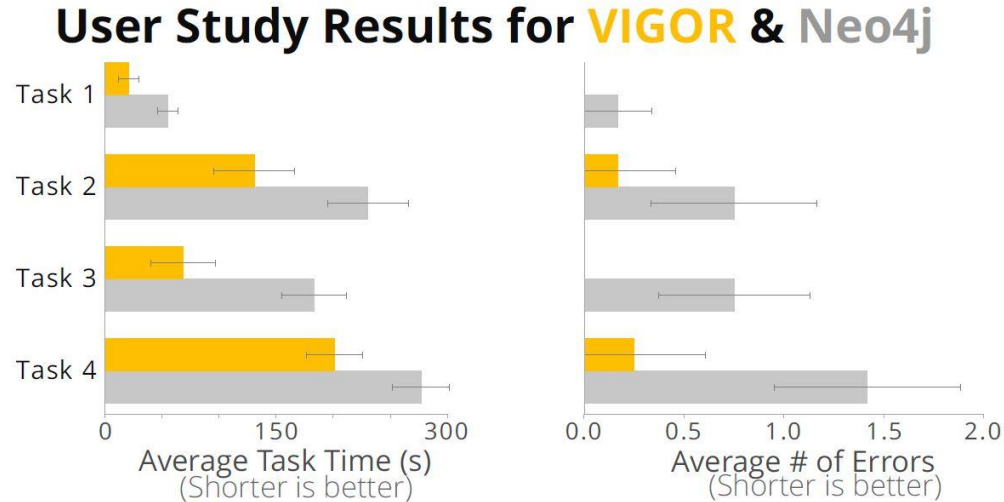


Experimental results (contd.)

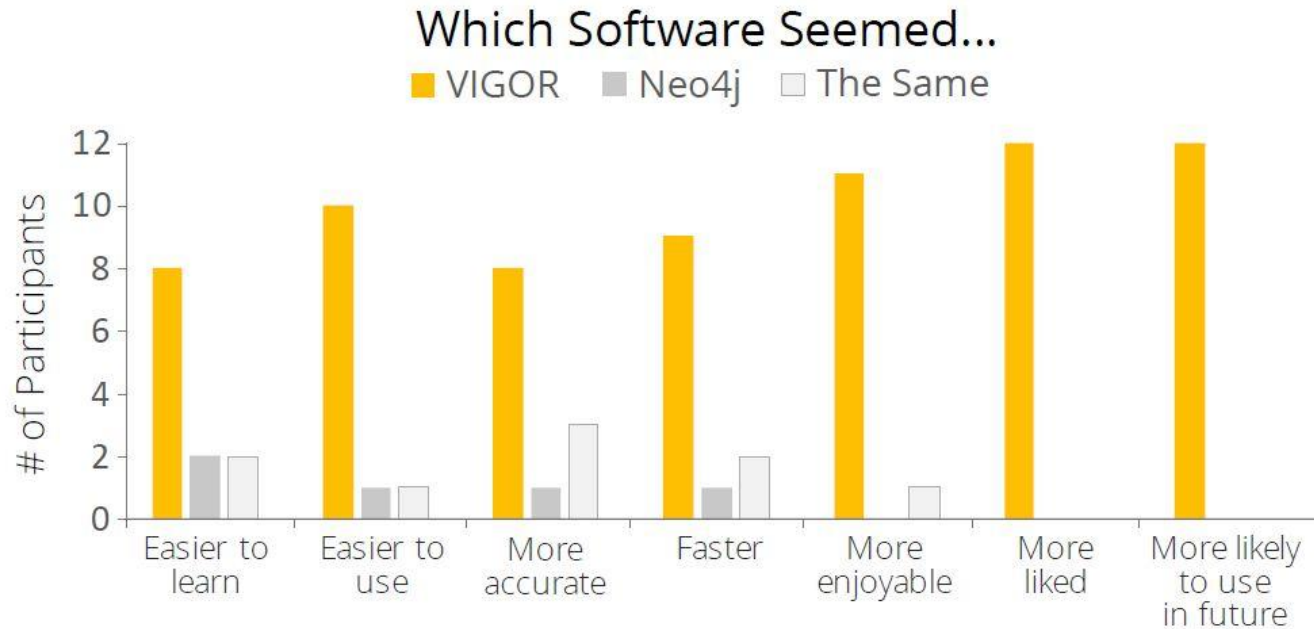
- ❑ Confounding factors:
 - ✓ Software (VIGOR or Neo4j)
 - ✓ Complexity of the task
 - ✓ Software Order (VIGOR or Neo4j going first)

- ❑ ANOVA results show only varying the software produced significant improvement.

Experimental results (contd.)



Experimental results (contd.)

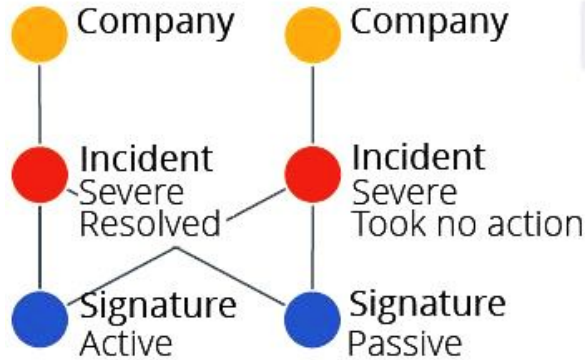


Slides adapted from Robert Pienta et al. @GaTech with the authors' permission

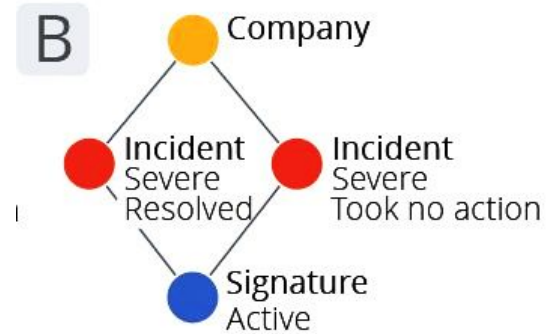
User Study 2: Cyber Security

- ❑ Network: Cyber Security network with 7,651 nodes and 384,182 edges
- ❑ Nodes connect clients of Symantec with Security incidents
 - ✓ Active signature
 - ✓ Passive signature
- ❑ Dataset contains > 11,000 incidents
- ❑ Participants: 3 Cyber Security experts from Symantec
- ❑ Number of queries: 2

Experimental results

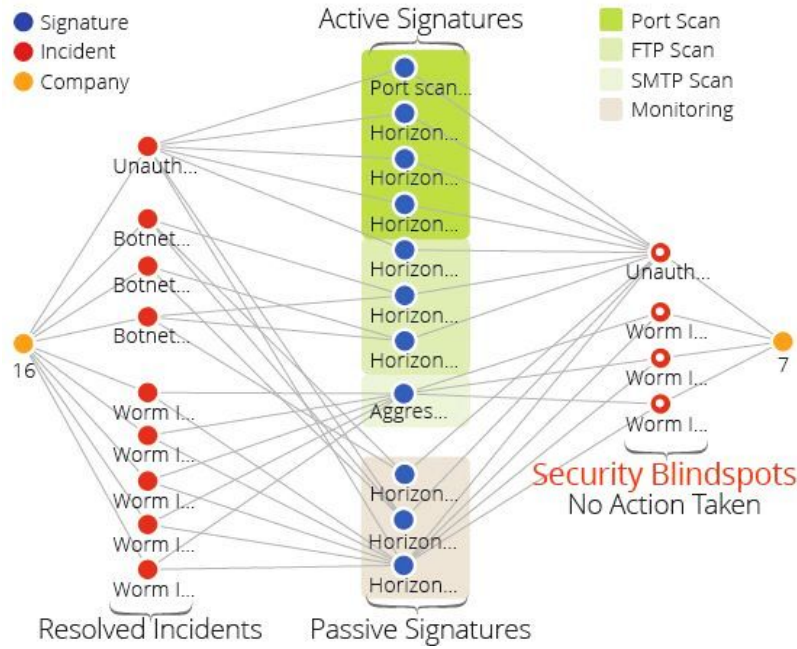


Query 1: Compare two companies with at least one active and one passive signature where one took actions to every threat detected and one did not



Query 2: Find companies that are inconsistent in taking actions when a threat is detected

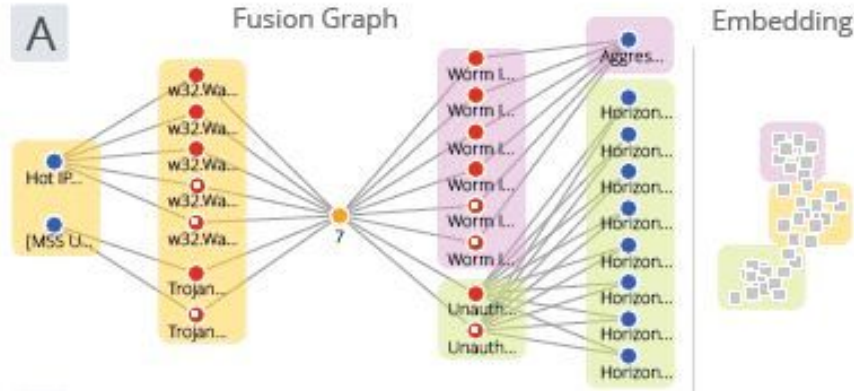
Experimental results (contd.)



Visualization of Query 1 results:

Security blindspots for Company "7" is detected using Fusion Graph

Experimental results (contd.)



Visualization of Query 2 results:

A company with inconsistent behavior when an incident is reported

Pros/Cons

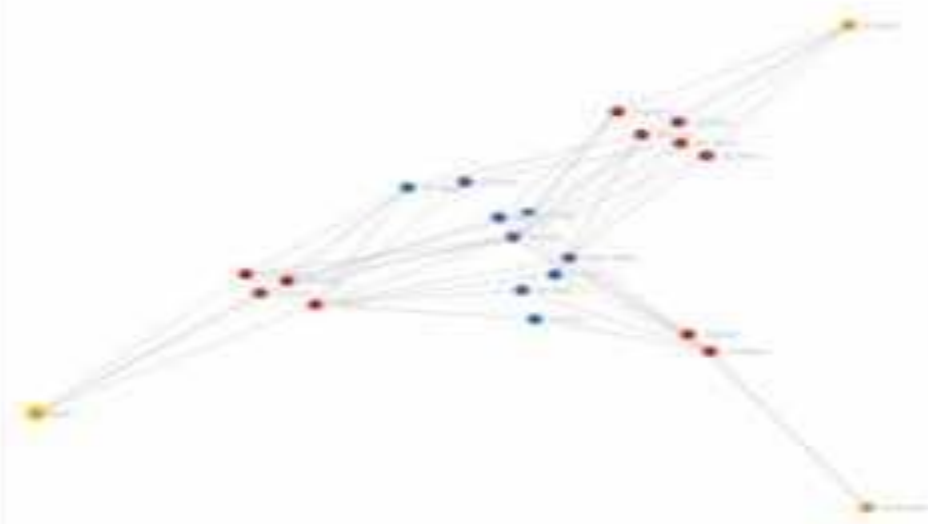
Pros:

- ❑ Exemplar-based interaction technique
- ❑ Feature-aware subgraph result summarization

Cons:

- ❑ User study only with few professional participants
- ❑ Users should be familiar with Cypher QL
- ❑ No arguments are provided for used structural features
- ❑ Insufficient analysis on clustering quality
- ❑ No evaluation on query creation and refinement
- ❑ Some information loss in the embedding process
- ❑ No information about the scope of the graph results that can be visualized (directed?, weighted?)
- ❑ User cannot filter based on the features in the feature explorer view

Feature Graph (100 features, 100 samples)



Expected Results



Thank You!

