Project Konigsberg

a Graph AI

Denis Vrdoljak

Managing Director, BDSG

Data Science Instructor, UC Berkeley

Gunnar Kleemann, Ph.D.

Senior Data Scientist, BDSG

Data Science Instructor, UC Berkeley

Project Konigsberg

a Graph AI

THE ELEVATOR PITCH

Project Konigsberg

a Graph AI

Gunnar Kleemann

Denis Vrdoljak

Project Konigsberg

a Graph AI

BDSG

Berkeley Data Science Group

Founded by Berkeley Data Science Instructors to bring products and services to market, and to bridge the gap between commercial and academic projects.

Project Konigsberg

a Graph AI

Road Map

Background Story
The Problem
Relational Data and ML
A New Approach
Examples
Applications
Future Development

Project Konigsberg

a Graph AI

Background Story

The Problem
Relational Data and ML
A New Approach
Examples
Applications
Future Development

Evolution of Project Konigsberg

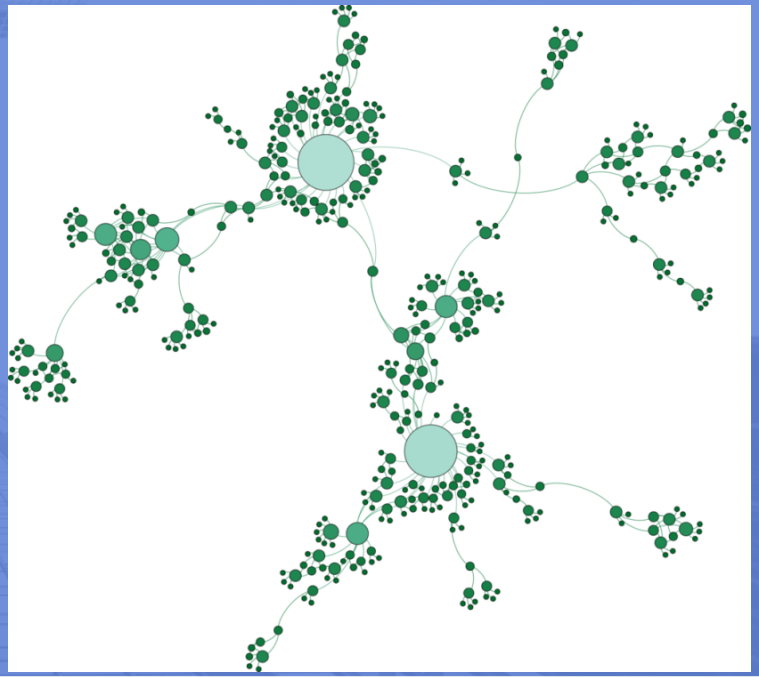
Denis Vrdoljak | Gunnar Kleemann | Danny Wudka

BioRevs: Biotech IPO Predictive Analytics Tool
26 April 2016
UC Berkeley School of Information

Researcher Graphs & Network Analysis
22 August 2016
Dojo Bali

BioRevs IPO Predictive Analytics, with Graph Analysis
17 January 2017
Data Day Austin

Project Konigsberg: a GraphAI
17 June 2017
Graph Day SF



24,000,000
Full PubMed Dataset (all lit)

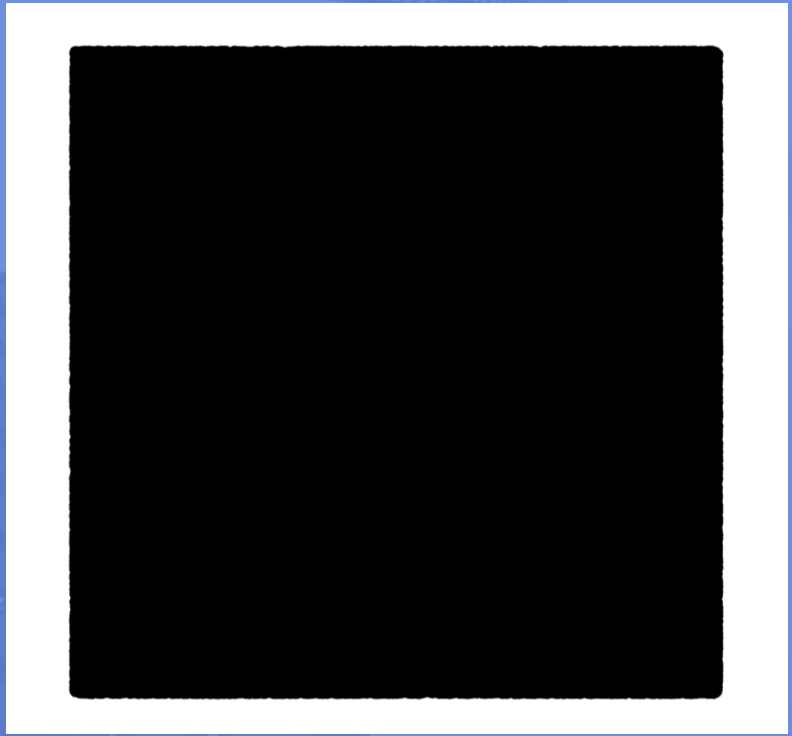
—

2,900,000 (~1.5 Million Open Source)
PubMed Central (BioTech -open access subset)

—

1,219,850
Total Entries, Condensed, after Cleaning/Pre-Processing

How to Parallelize for Scaling?



Our initial result

350,000 nodes
(1/10th of our dataset)

That’s an Actual Plot, not an Error

51,700
Job Skills on LinkedIn

—

20,000 **(Jobs /major metro area /job title)**

—

2,000,000
Relationships to score, per search, after cleaning data

Konigsberg: a GraphAI

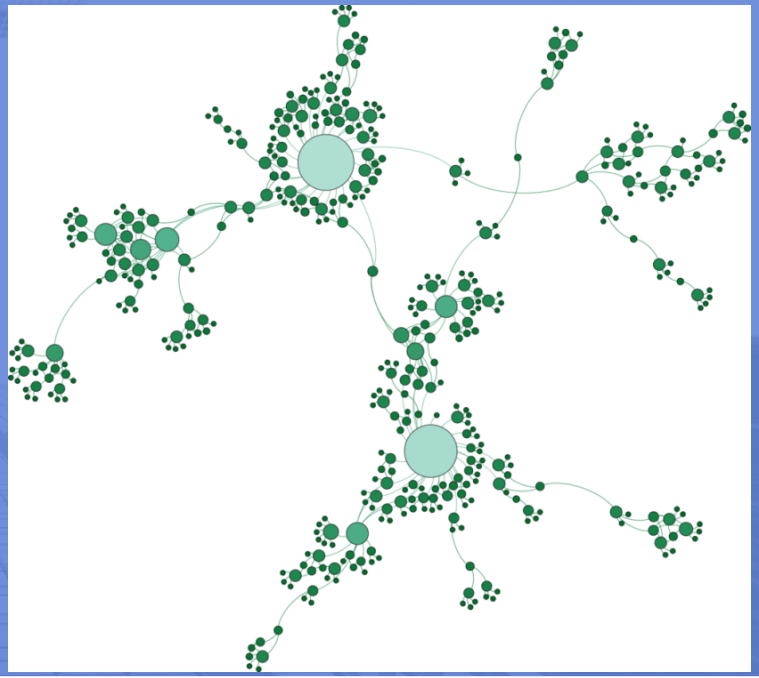
Common Pain Points

Graph Data in Traditional ML Models/Tools

Capturing Context, other Extra Dimensions

Higher Dimensional Relationships in TDIDF

O(n) Cost when Computing Graph Metrics



Project Konigsberg

a Graph AI

Background Story

The Problem

Relational Data and ML
A New Approach
Examples
Applications
Future Development

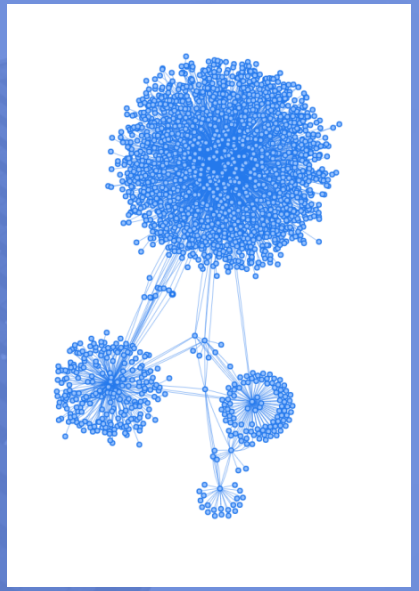
The Problem: Key Points

Characterizing and Representing Topologies
Predicting Missing Edges
DeConvoluting Overlapping Nodes (e.g., key="John Smith")
Finding a Suitable Cost Function/Measure for ML
Identifying Outliers
Classifying Connection Strengths

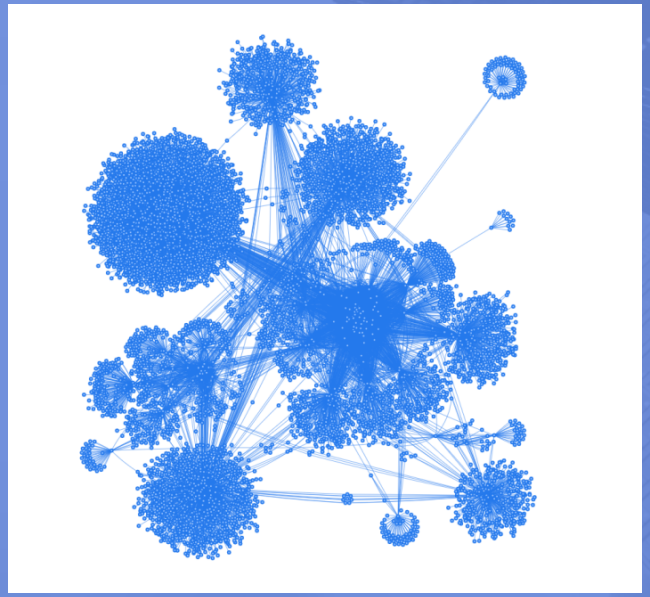
…and Doing All this For Varying Graph Topologies/Datasets

Characterizing Network Topologies

Scott W Emmons

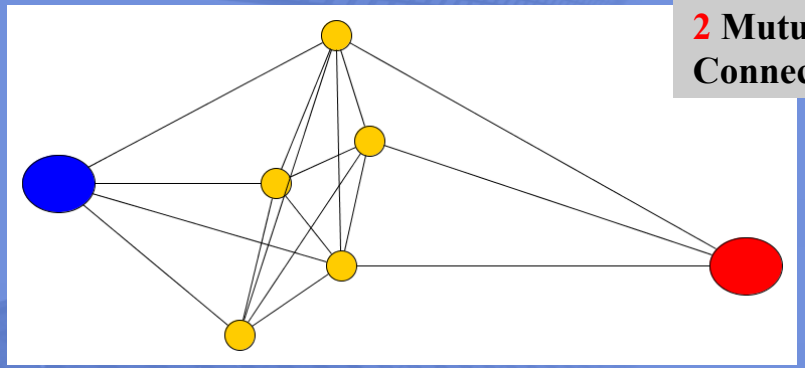


Coleen Murphy

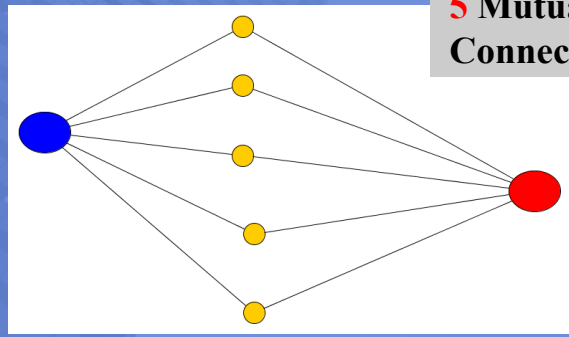


Is One of These an Outlier?

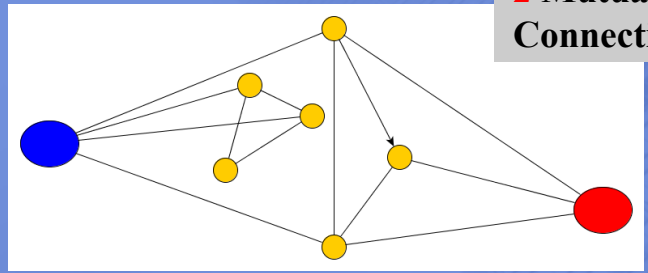
Which Ones Are Connected?



2 Mutual Connections



5 Mutual Connections



2 Mutual Connections

Project Konigsberg

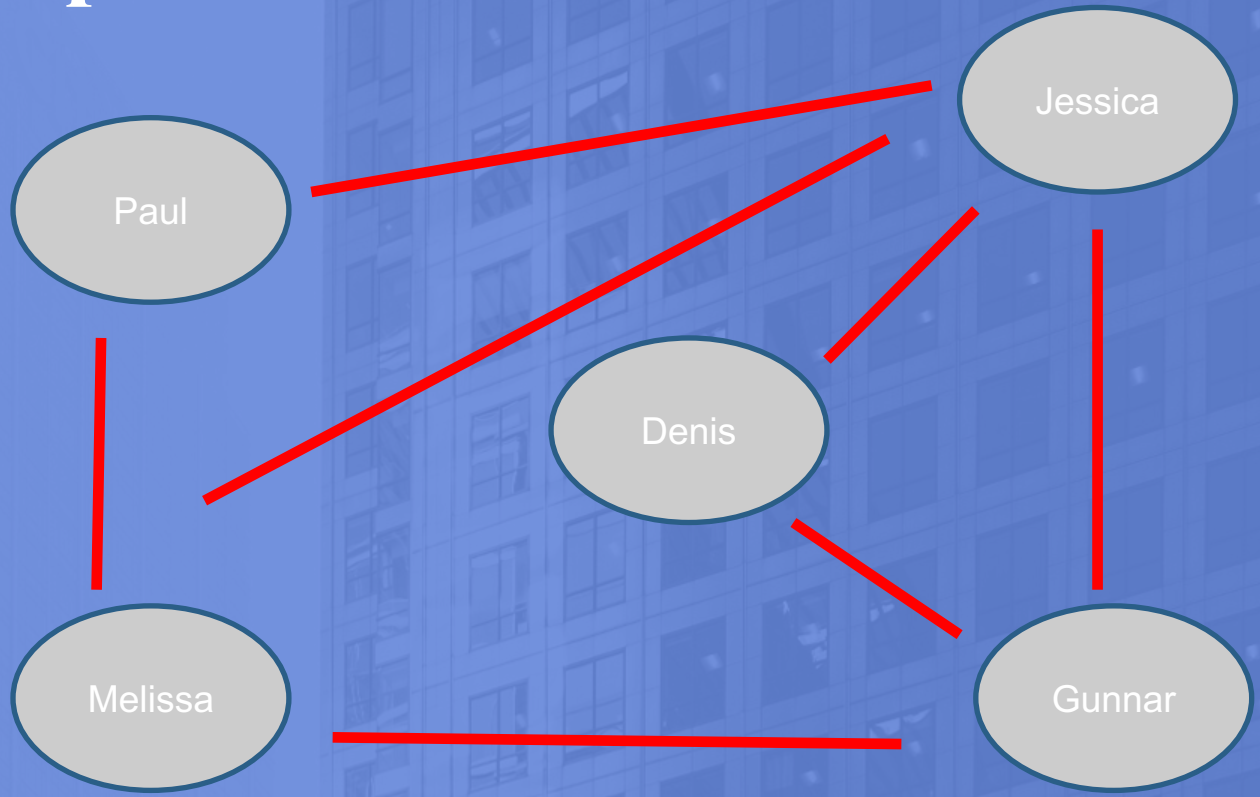
a Graph AI

Background Story
The Problem

Relational Data and ML

A New Approach
Examples
Applications
Future Development

A Simple Social Network



Adjacency Matrix

	Paul	Denis	Melissa	Jessica	Gunnar
Paul	0	0	1	1	0
Denis	0	0	0	1	1
Melissa	1	0	0	1	1
Jessica	1	1	1	0	1
Gunnar	0	1	1	1	0

Shortest Path Matrix

	Paul	Denis	Melissa	Jessica	Gunnar
Paul	0	2	1	1	2
Denis	2	0	2	1	1
Melissa	1	2	0	1	1
Jessica	1	1	1	0	1
Gunnar	2	1	1	1	0

Shortest Path Matrix

Traditional Graph Metrics Don’t Represent the Information We Need

Project Konigsberg

a Graph AI

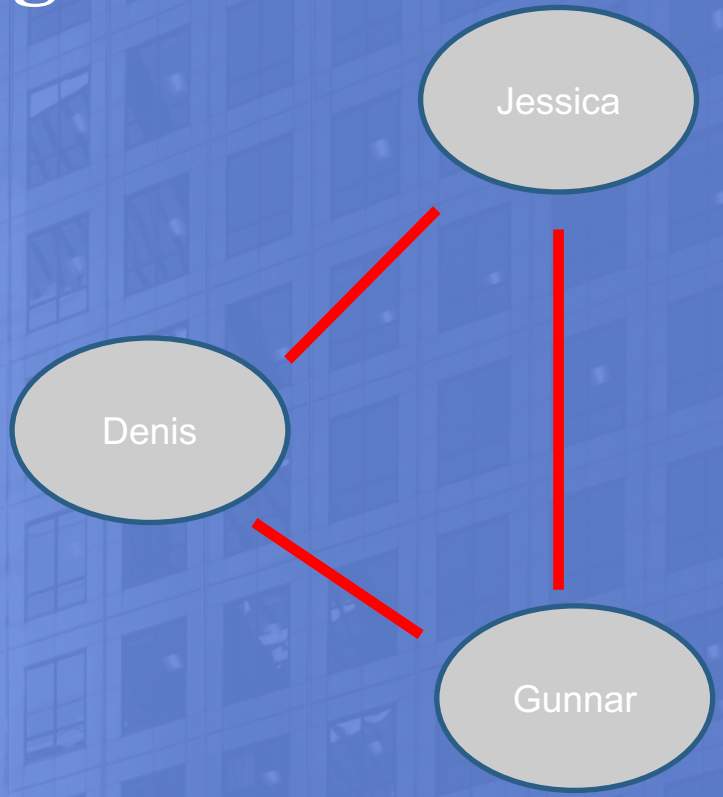
Background Story
The Problem
Relational Data and ML

A New Approach

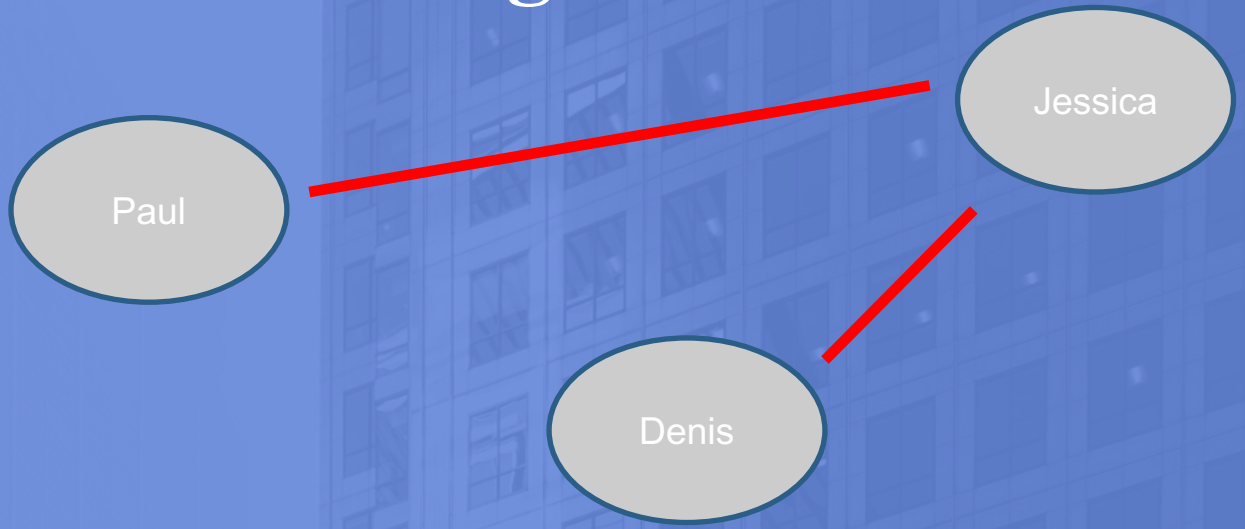
Examples
Applications
Future Development

First some Definitions

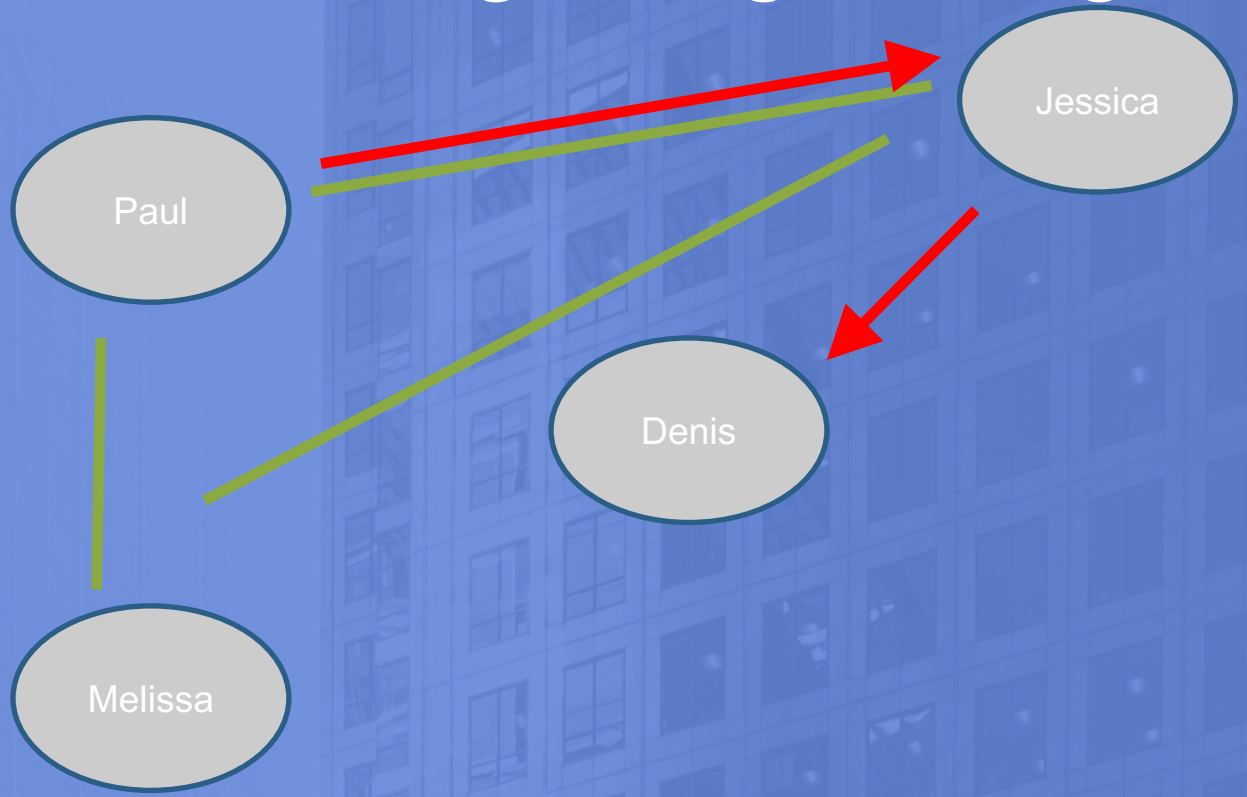
Definitions: Triangle



Definitions: Wedge

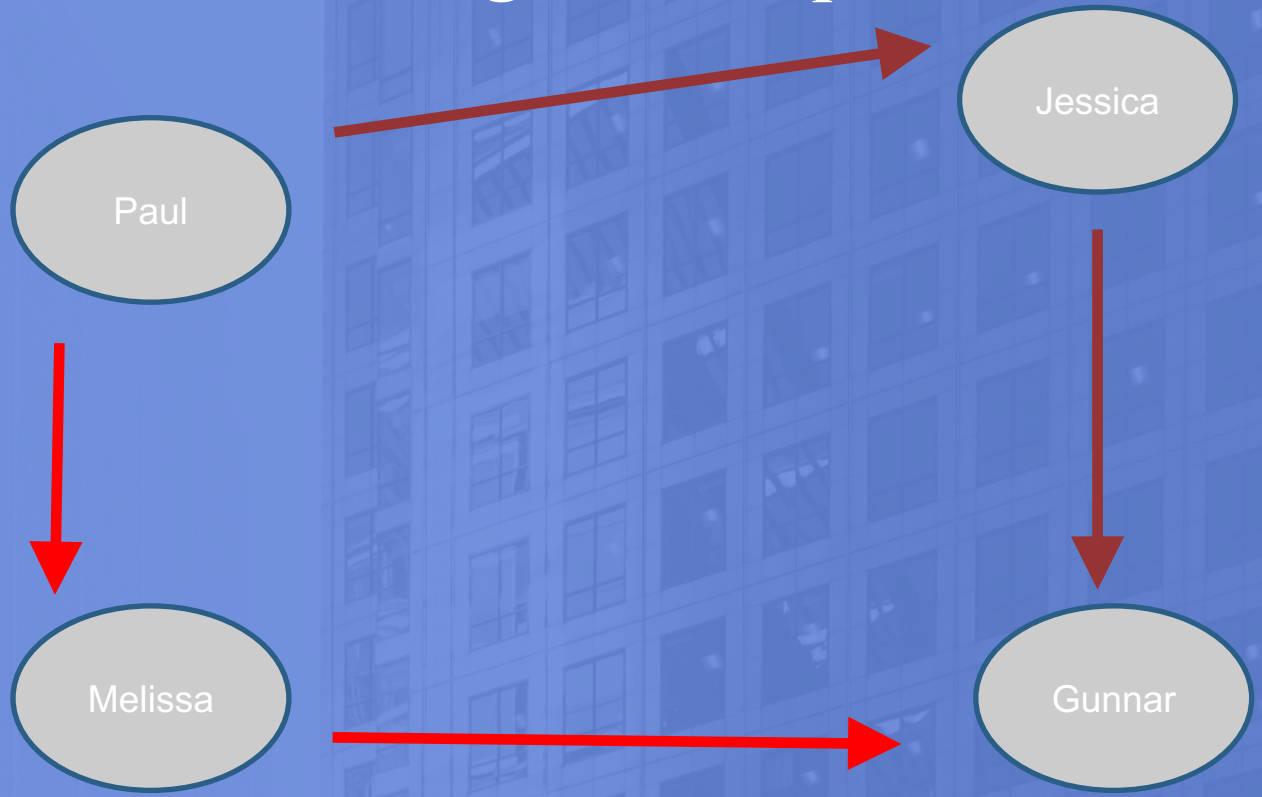


Definition: 1 Wedge through a Triangle Node

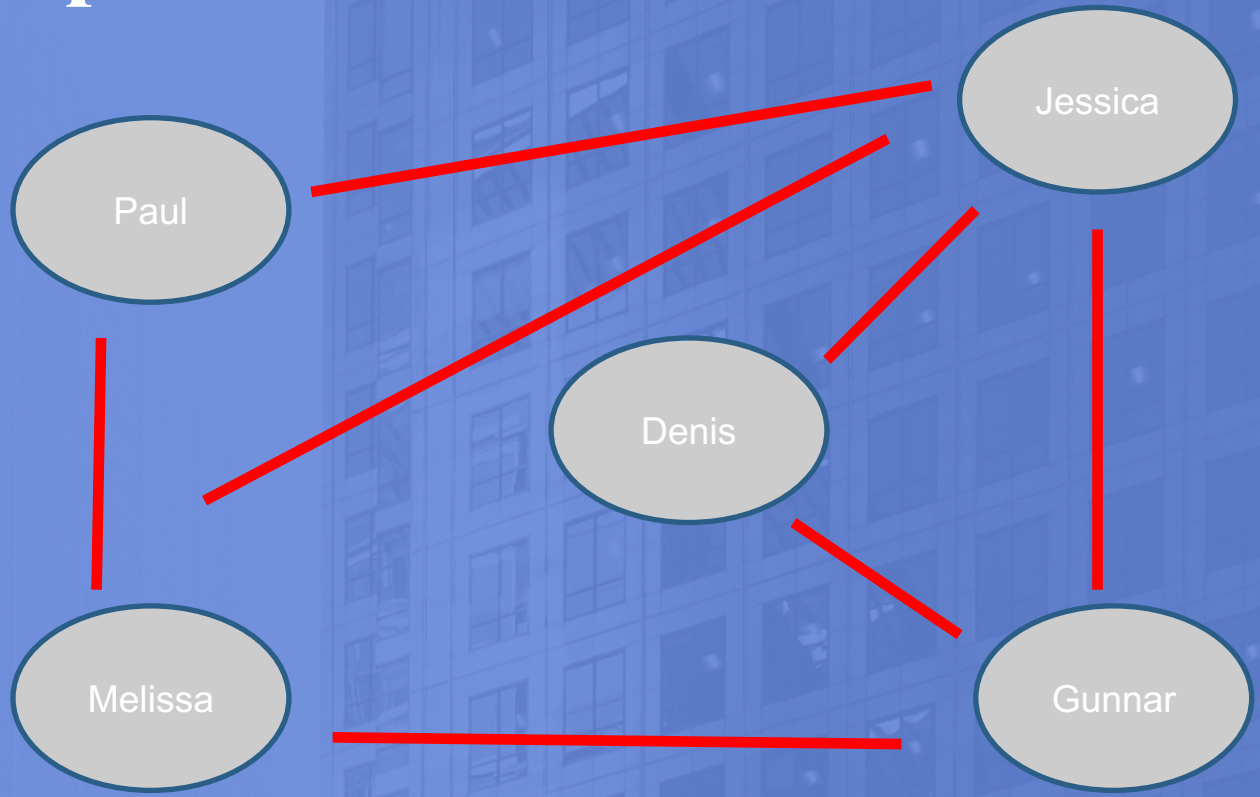


Definition: 2 Wedges via Triangle Nodes

Definition: 2 Wedges, 1 Unique End Node



A Simple Social Network



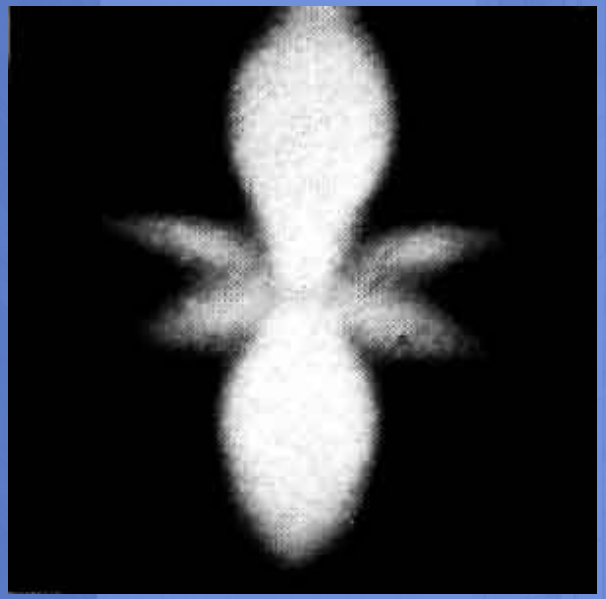
Some New Metrics

	Wedges	Triangles	Wedge through nodes found in (1,2,or 3) triangle	Wedges to unique nodes	Triangles with unique nodes
Paul	3	1	(3,0,0)	2	2
Denis	3	1	(3,0,0)	2	2
Melissa	2	2	(1,1,0)	1	3
Jessica	0	3	(0,0,0)	0	4
Gunnar	2	2	(1,1,0)	1	3

This Becomes a Probability Distribution

	Wedges	Triangles	Wedge through nodes found in (1,2,or 3) triangle	Wedges to unique nodes	Triangles with unique nodes
Paul	3	1	(3,0,0)	2	2
Denis	3	1	(3,0,0)	2	2
Melissa	2	2	(1,1,0)	1	3
Jessica	0	3	(0,0,0)	0	4
Gunnar	2	2	(1,1,0)	1	3

…to a MultiDimensional Probability Distribution



Project Konigsberg

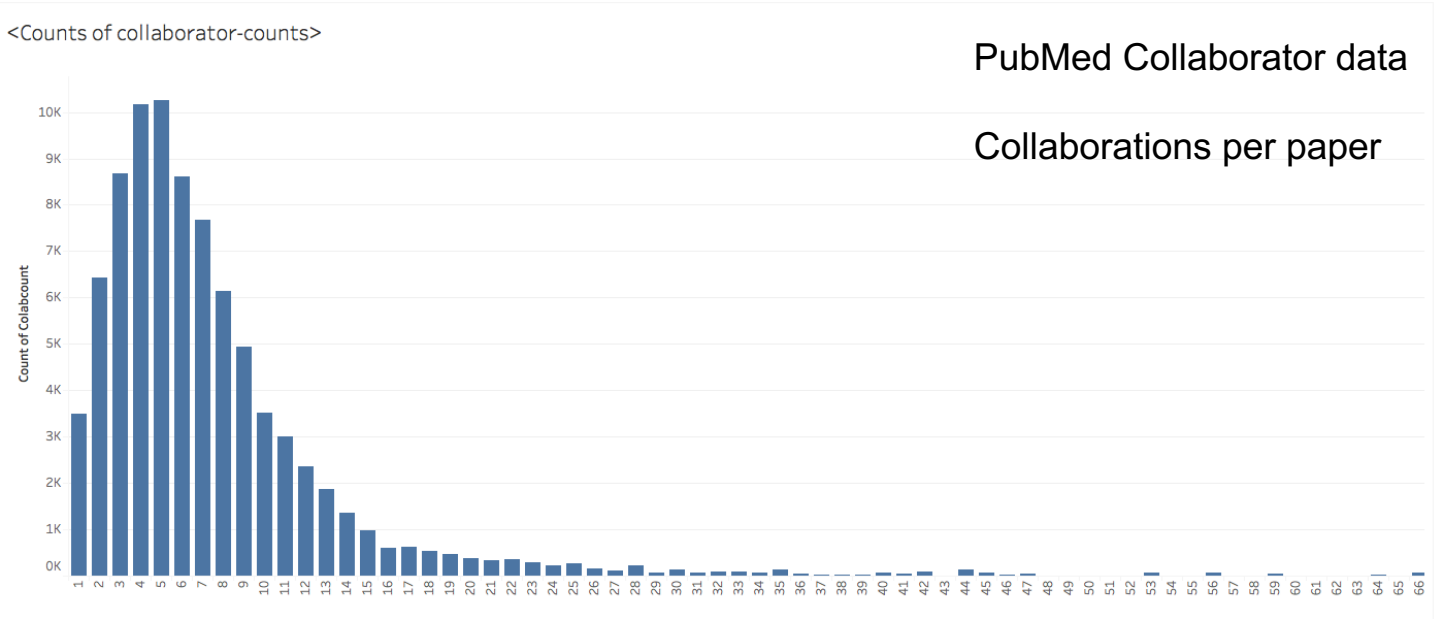
a Graph AI

Background Story
The Problem
Relational Data and ML
A New Approach

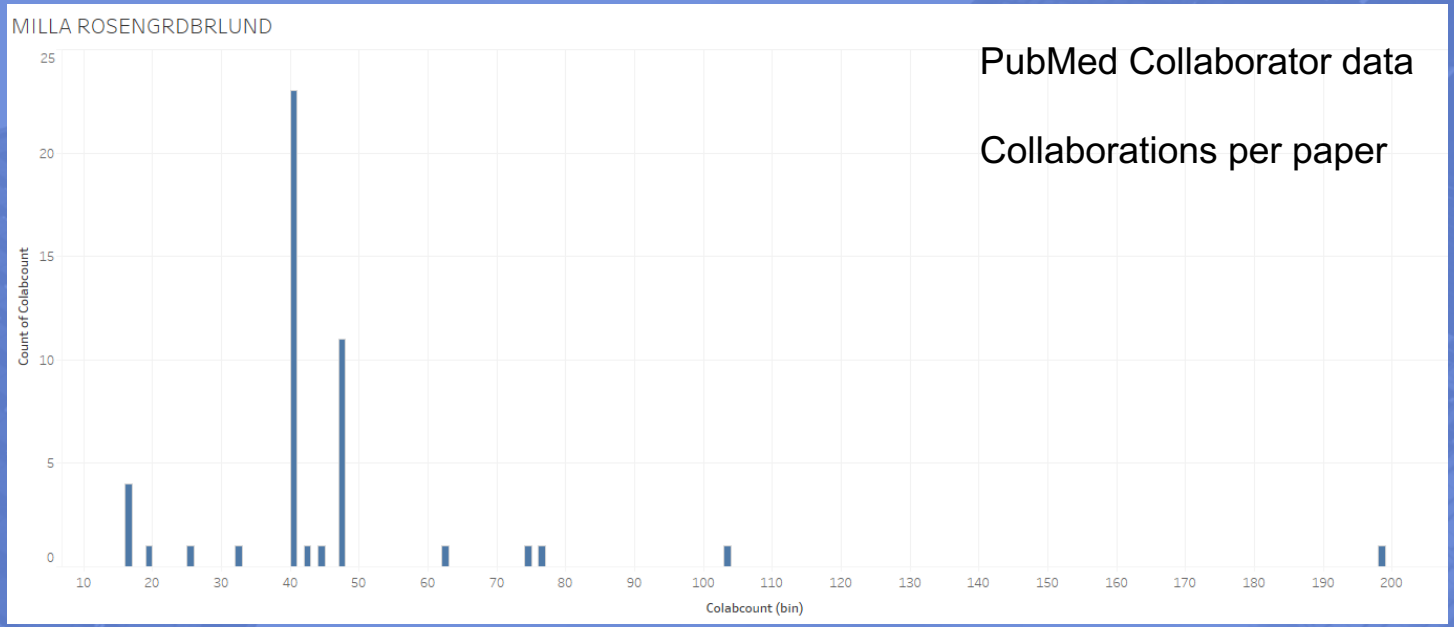
Examples

Applications
Future Development

Overall Frequency Distribution



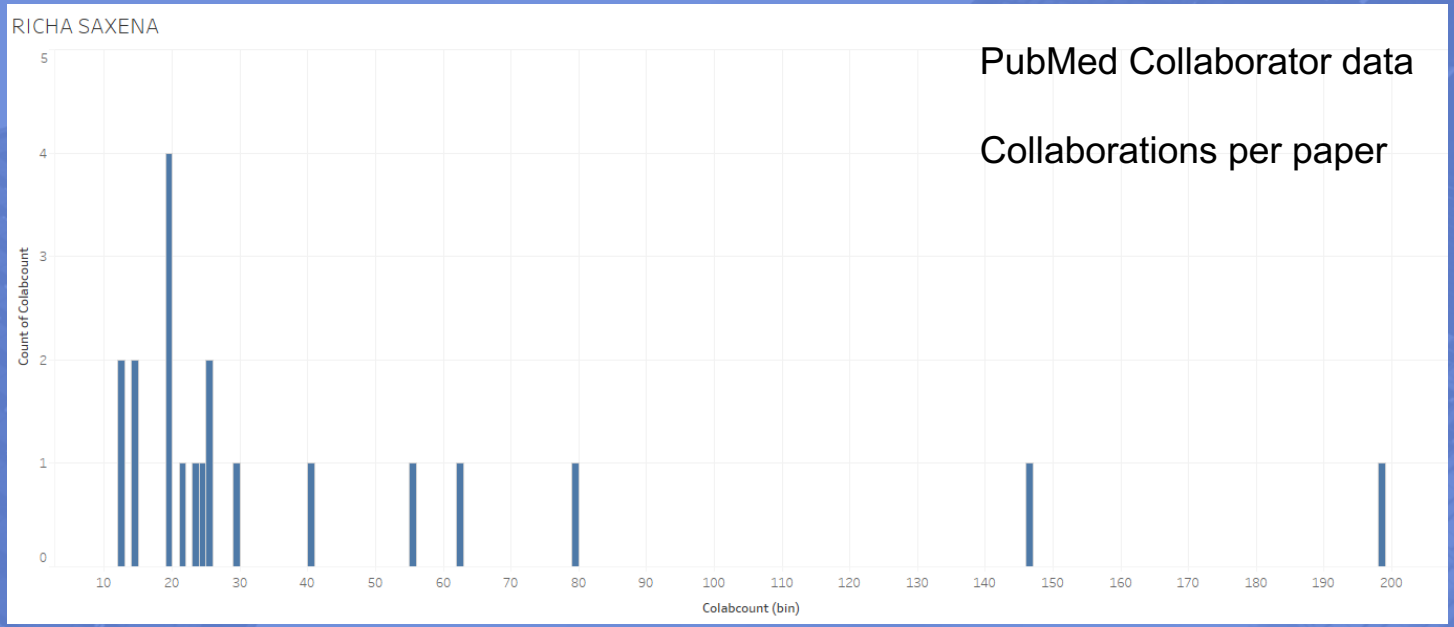
Milla Rosengrdbrlund Frequency Distribution



PubMed Collaborator data

Collaborations per paper

Richa Saxena Frequency Distribution

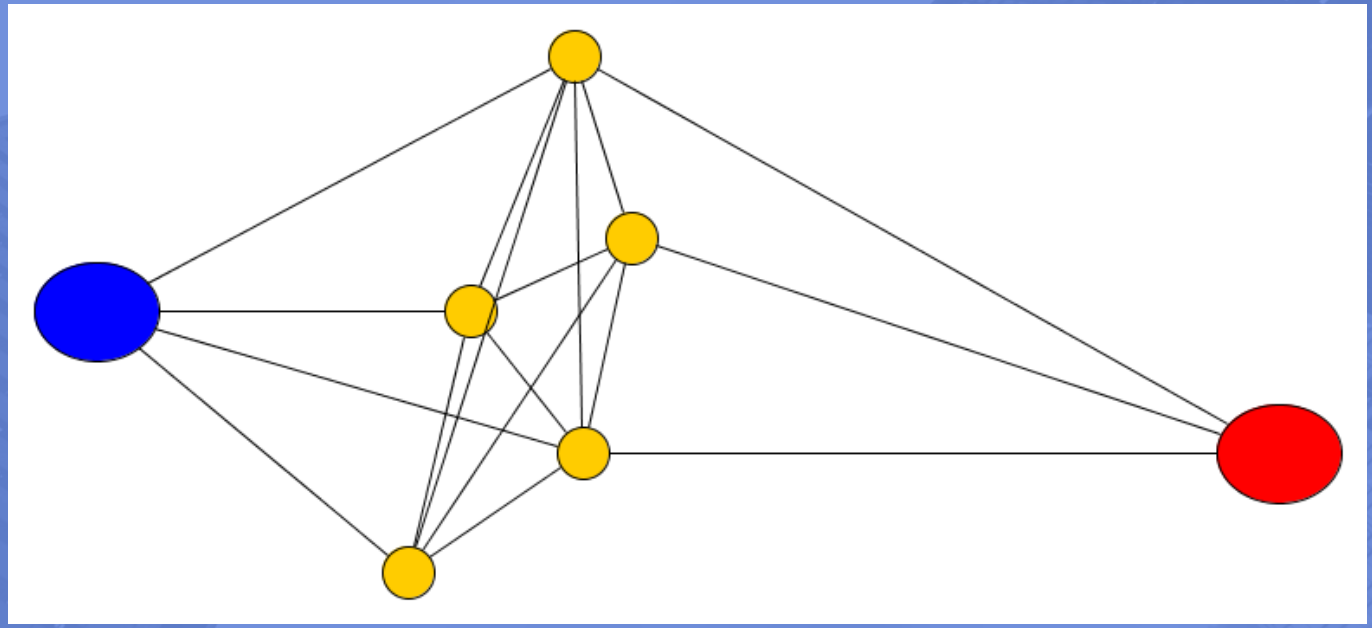


PubMed Collaborator data

Collaborations per paper

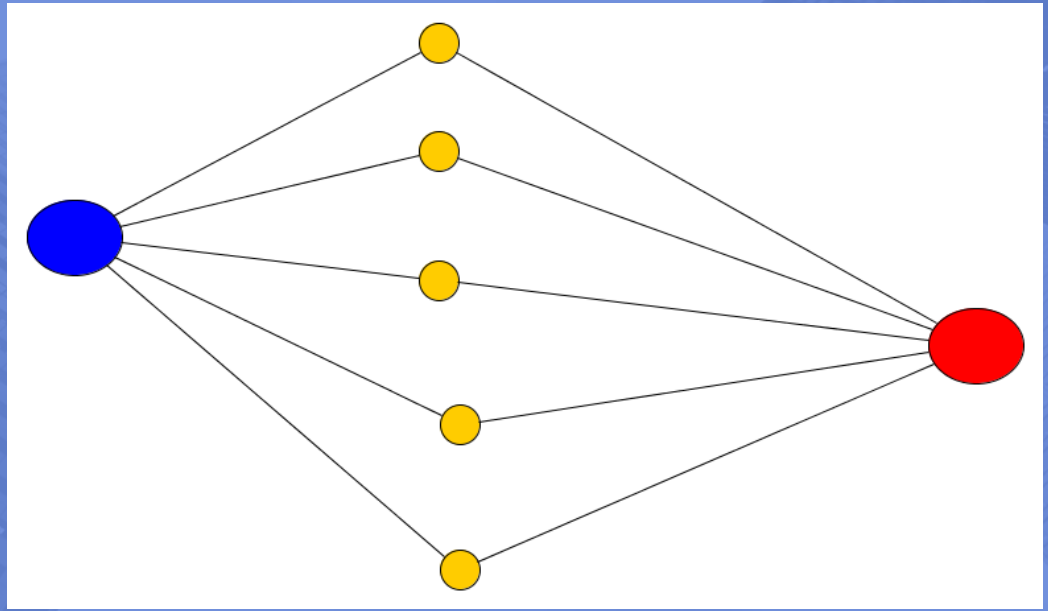
Predicting Edges

2 Mutual Connections



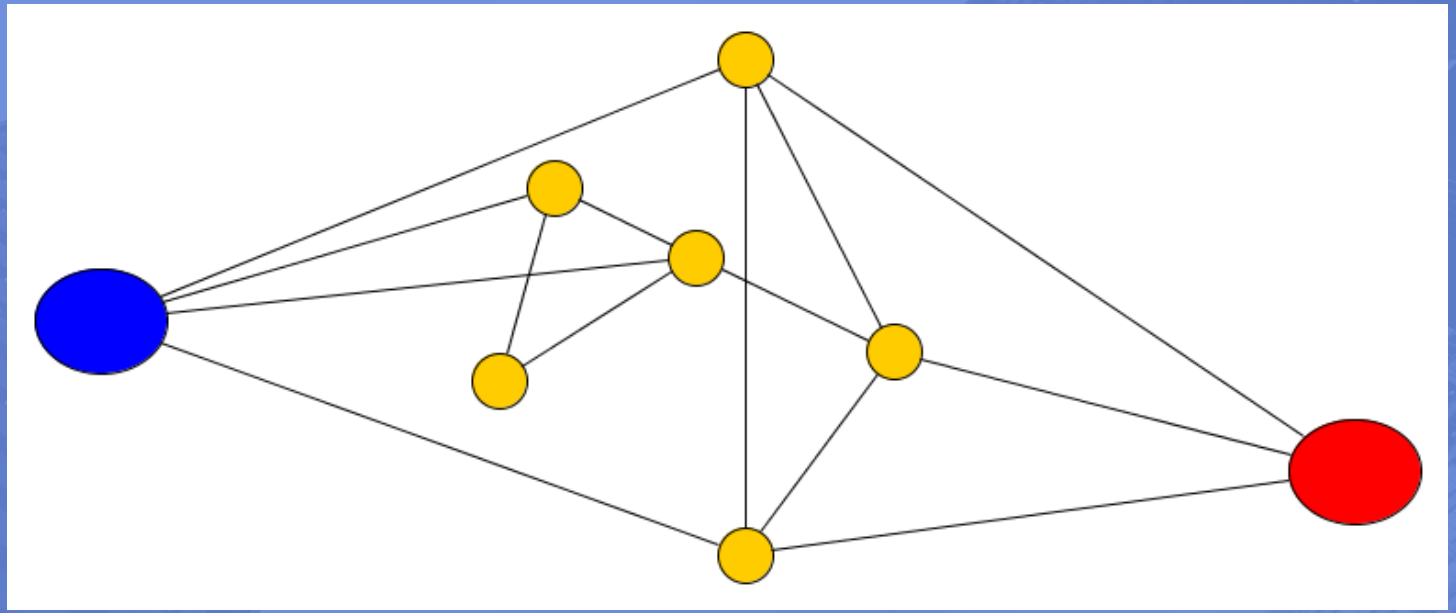
Predicting Edges

5 Mutual Connections



Predicting Edges

2 Mutual Connections



We detected top publishers and found scientist that we should know but do not.

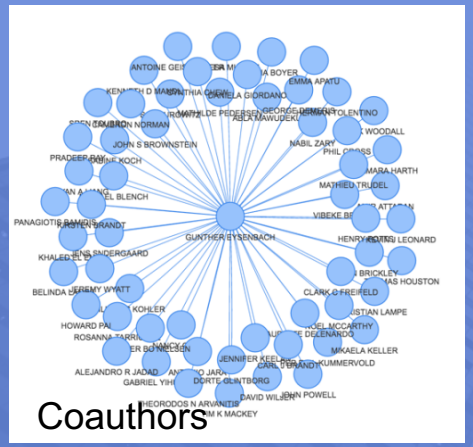
We identify the top publishers:

*** Gunther Eysenbach**
*** Prof. Hoongkun Fun**
*** Seik Weng Ng**

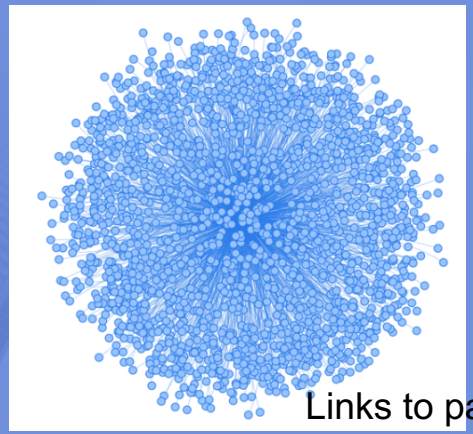
*** we need a better way to identify top scientists when they have common names.**

```
neo4j-sh (?)$ MATCH (n)-[r:AUTHORED]->(m)
> RETURN n, COUNT(r)
> ORDER BY COUNT(r) DESC
> LIMIT 10;
+-------------------------------------------------+
| n                                    | COUNT(r) |
+-------------------------------------------------+
| Node[487153]{name:"THE  "}           | 2671     |
| Node[865318]{name:"GUNTHER EYSENBACH"} | 2055     |
| Node[78005]{name:"HOONGKUN FUN"}     | 1329     |
| Node[1520956]{name:"SEIK WENG NG"}   | 1233     |
| Node[12886]{name:"WEI WANG"}         | 1203     |
| Node[33084]{name:"WEI ZHANG"}        | 1029     |
| Node[148]{name:"YAN LI"}             | 829      |
| Node[6653]{name:"WEI LI"}            | 778      |
| Node[15555]{name:"JING WANG"}        | 737      |
| Node[20702]{name:"LI ZHANG"}         | 670      |
+-------------------------------------------------+
10 rows
24526 ms
```

Few Co-Authors, Many Papers



Coauthors



Links to papers

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”.

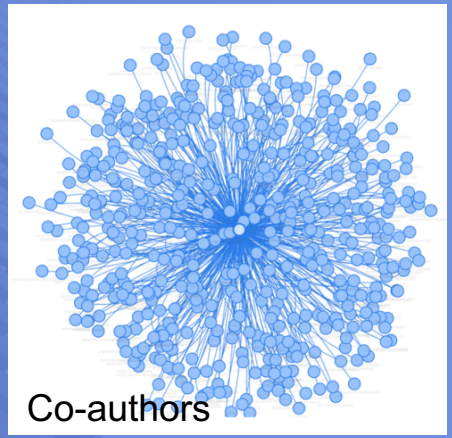
Founder of an academic field!

- association between search engine queries and influenza incidence,
- He coined the terms **"infoveillance" and "infodemiology"** for these kinds of approaches.

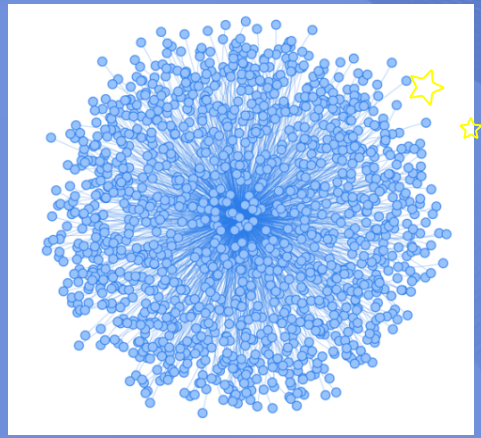
Source: http://ehealthinnovation.org/people/gunther-eysenbach/
Source: https://en.wikipedia.org/wiki/Gunther_Eysenbach

Prof. Kun has a pattern closer to what we expected

Many collaborators and many publications



Co-authors

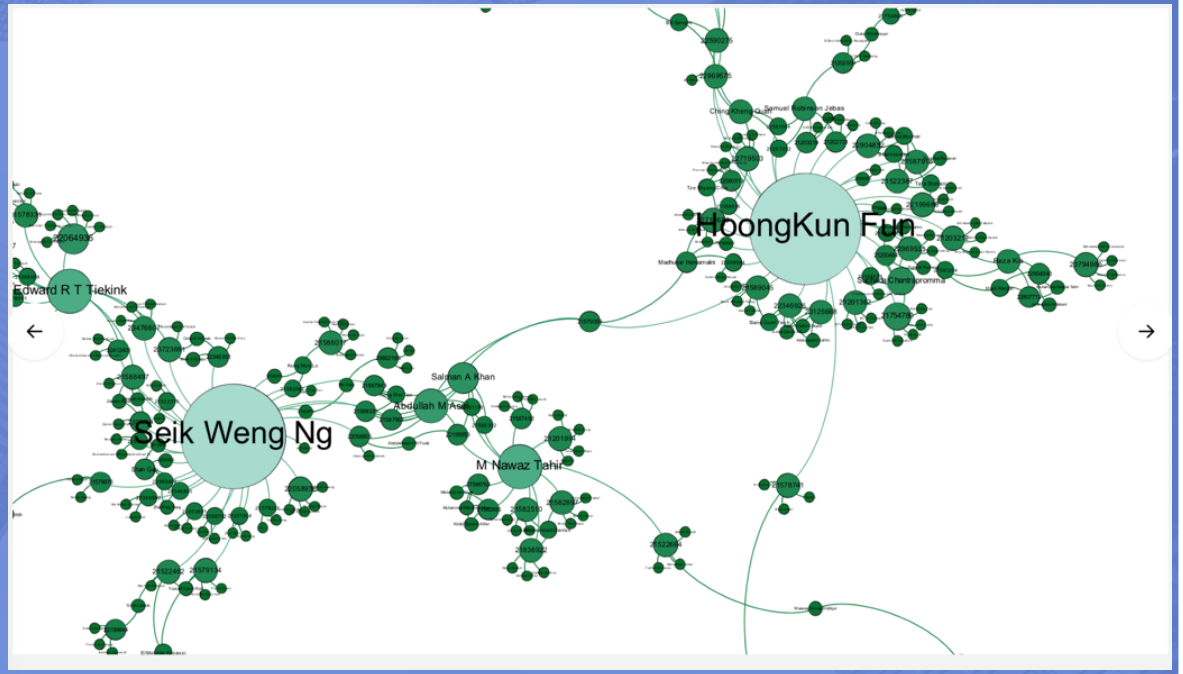


Papers

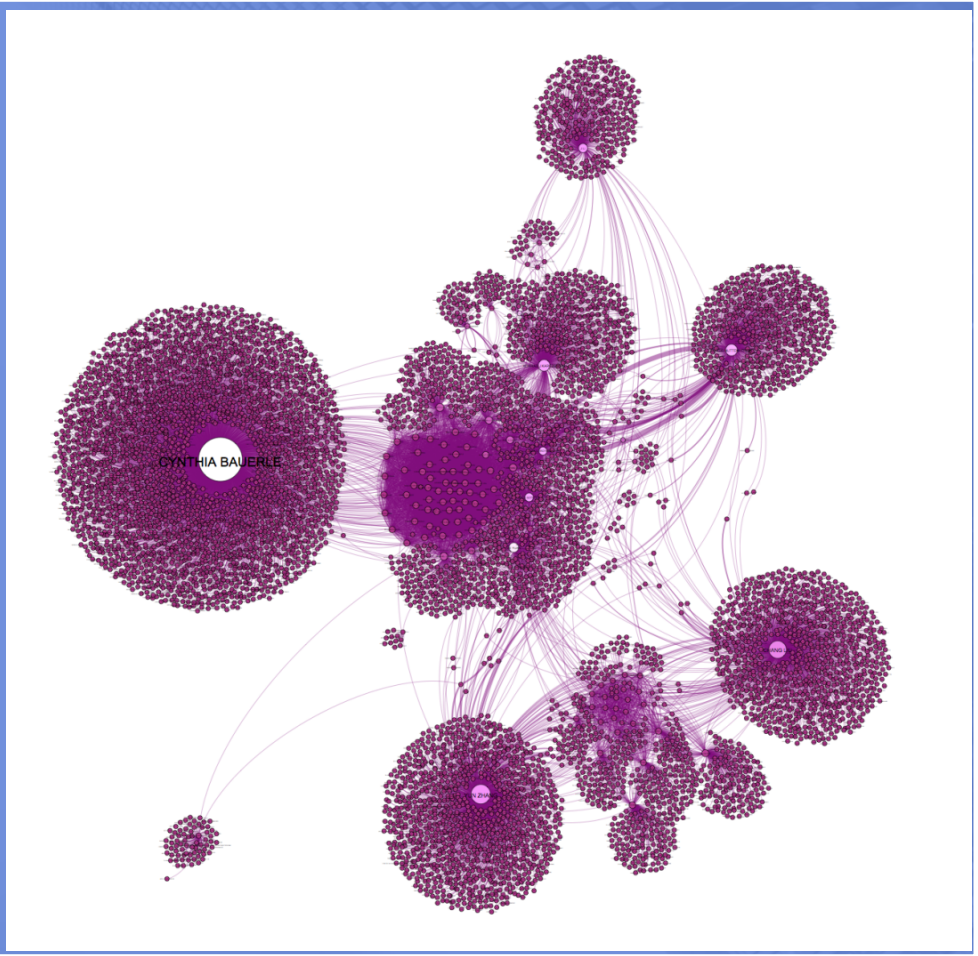
Fun Hoong Kun

Pubmed Dataset

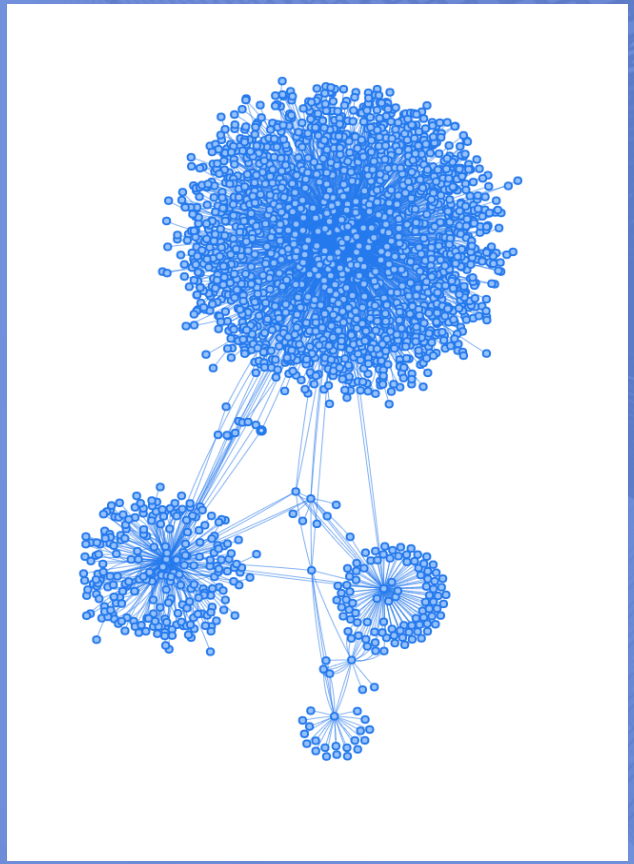
Graphing small samples



Coleen Murphy's 2nd Order Connections

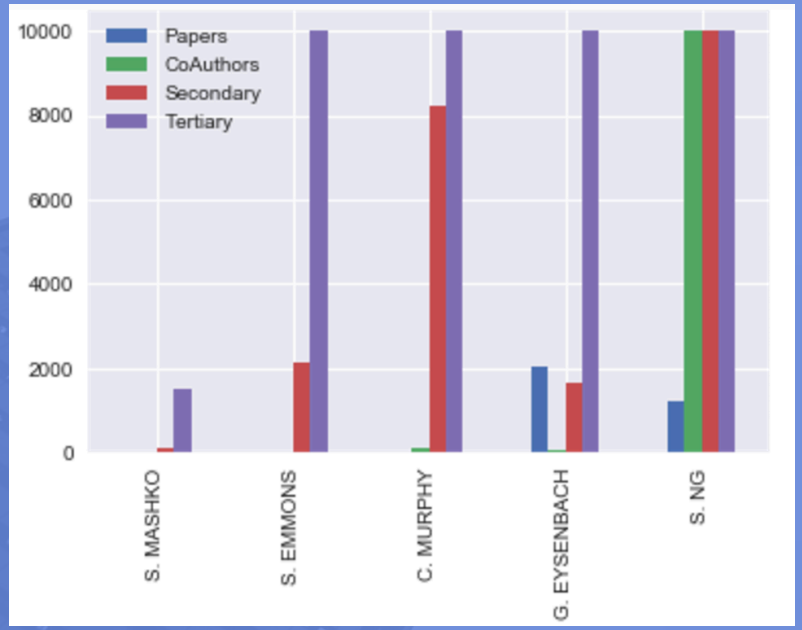


Scott Emmons's
2nd Order Connections

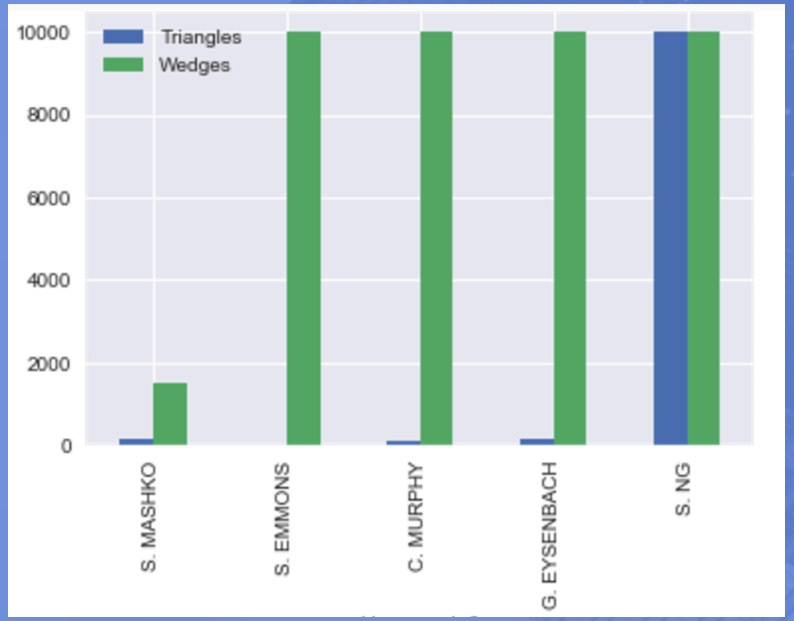


Graph traversal shows different collaboration patterns

Paper count, Primary, secondary and tertiary



Triangles vs wedges



Project Konigsberg

a Graph AI

Background Story
The Problem
Relational Data and ML
A New Approach
Examples

Applications

Future Development

Applications

UC Berkeley School of Information

ParlourBoard

MarketGraph

BioPredictor

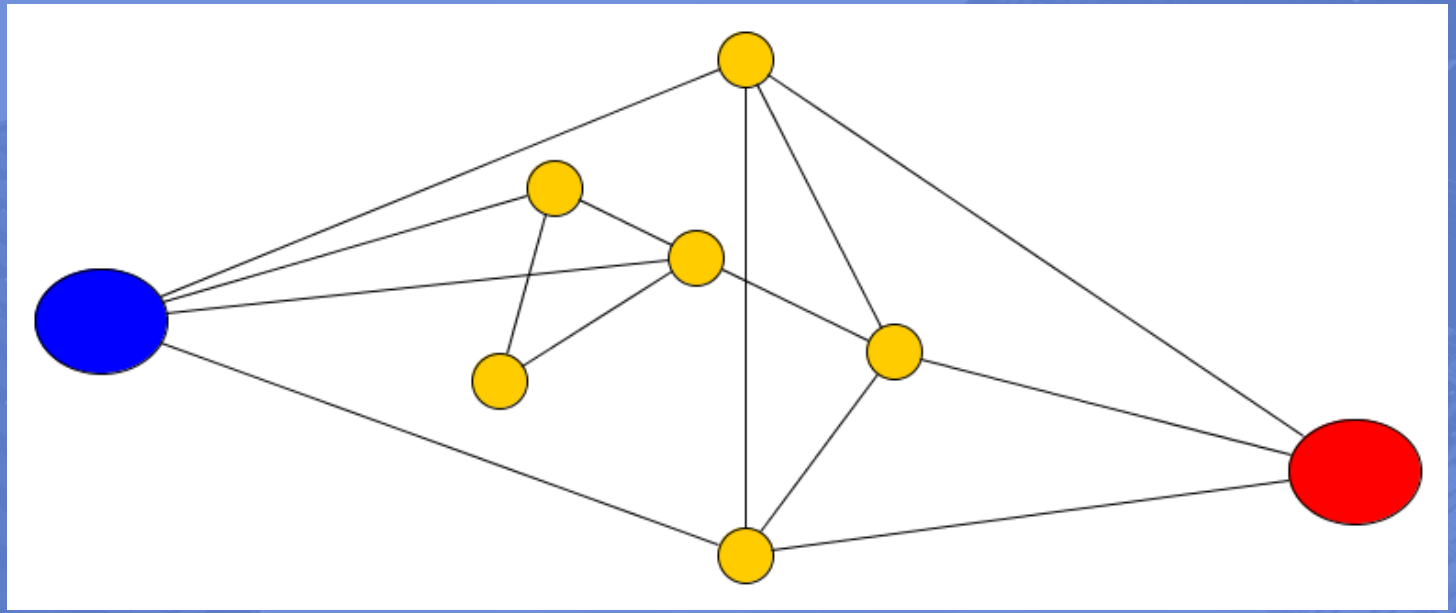
Project “Pigeon”

CAPITAL DATA AUSTIN

Marketing Analytics (Corp Customer)

LIVE DEMO

2 Mutual Connections



Project Konigsberg

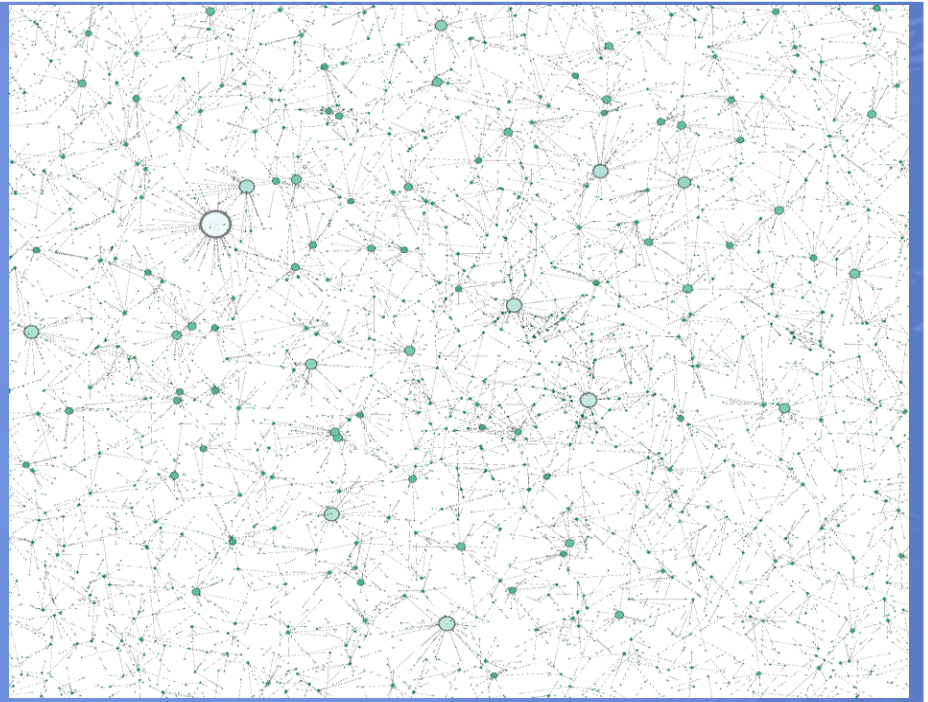
a Graph AI

Background Story
The Problem
Relational Data and ML
A New Approach
Examples
Applications

Future Development

Thank You

Any Questions?



Denis Vrdoljak

denis@bds.group

Gunnar Kleemann, Ph.D.

gunnar@bds.group

http://www.bds.group