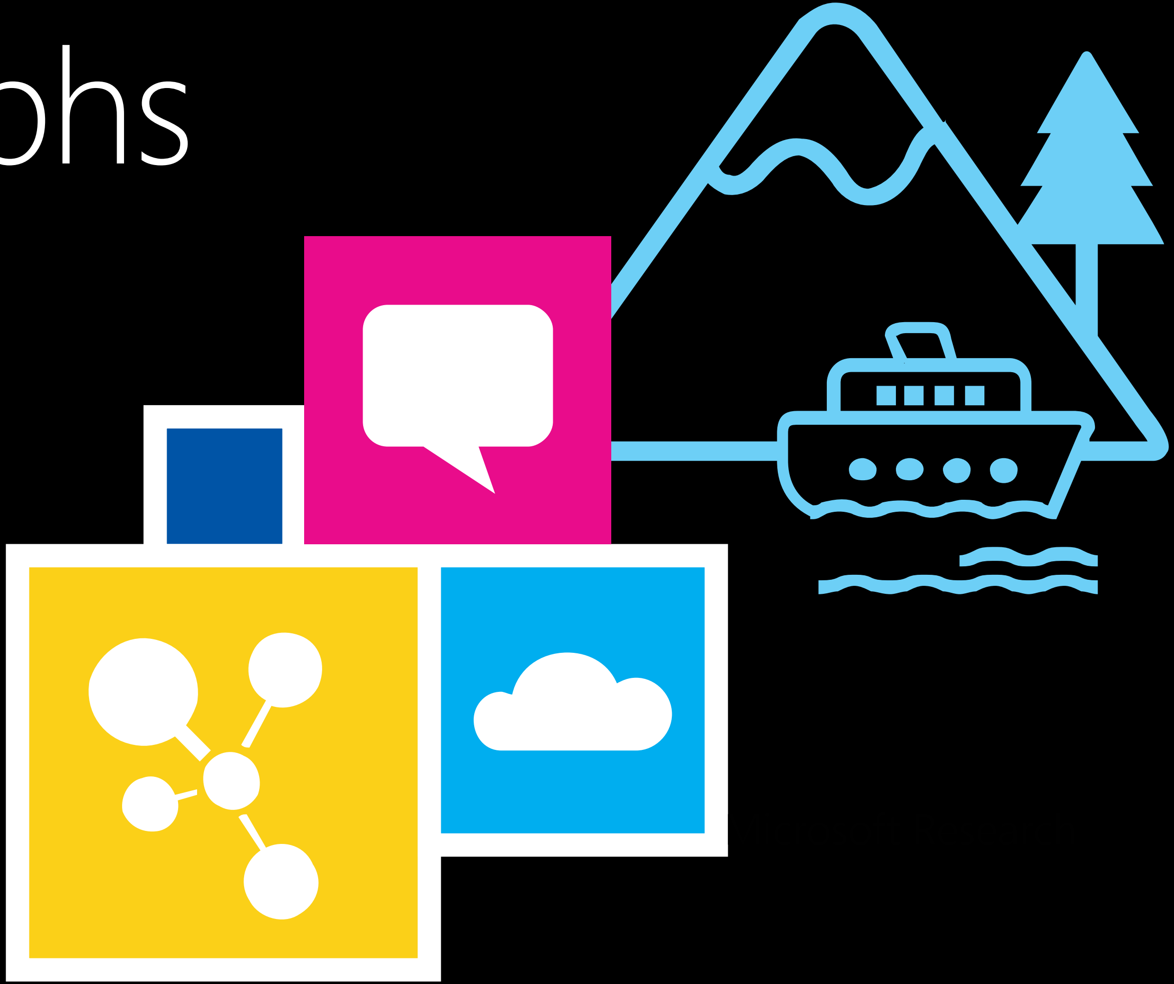
Microsoft

Interacting with Graphs

Lise Getoor
University of Maryland



BIG Data is not flat

©2004-2013 lonnitaylor

Data is multi-modal, multi-relational, spatio-temporal, multi-media

shorthand: **Graph Data**

NEED: ML[*] for Graphs

[*]: Machine Learning

ML for Graphs

Pattern #1: Collective Classification

Pattern #2: Link Prediction

Pattern #3: Entity Resolution

ML for Graphs

Pattern #1: Collective Classification – inferring labels of nodes in graph

Pattern #2: Link Prediction

Pattern #3: Entity Resolution

ML for Graphs

Pattern #1: Collective Classification – inferring labels of nodes in graph

Pattern #2: Link Prediction – inferring the existence of edges in graph

Pattern #3: Entity Resolution

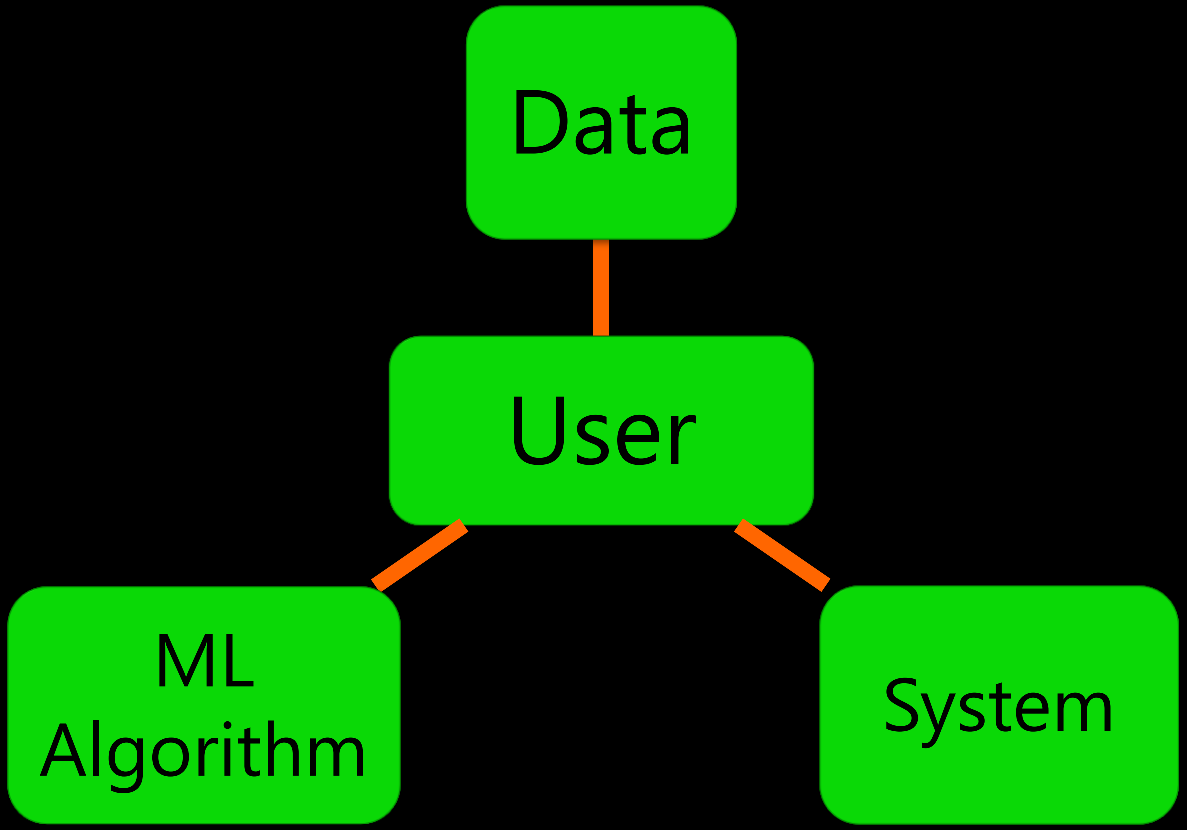
ML for Graphs

Pattern #1: Collective Classification – inferring labels of nodes in graph

Pattern #2: Link Prediction – inferring the existence of edges in graph

Pattern #3: Entity Resolution – clustering nodes that refer to the same underlying entity

What about Interaction?



Data

User

ML Algorithm

System

What’s different about graphs?

Unit of Interaction

Context

Comparison

What's different about graphs?

Unit of Interaction

Context

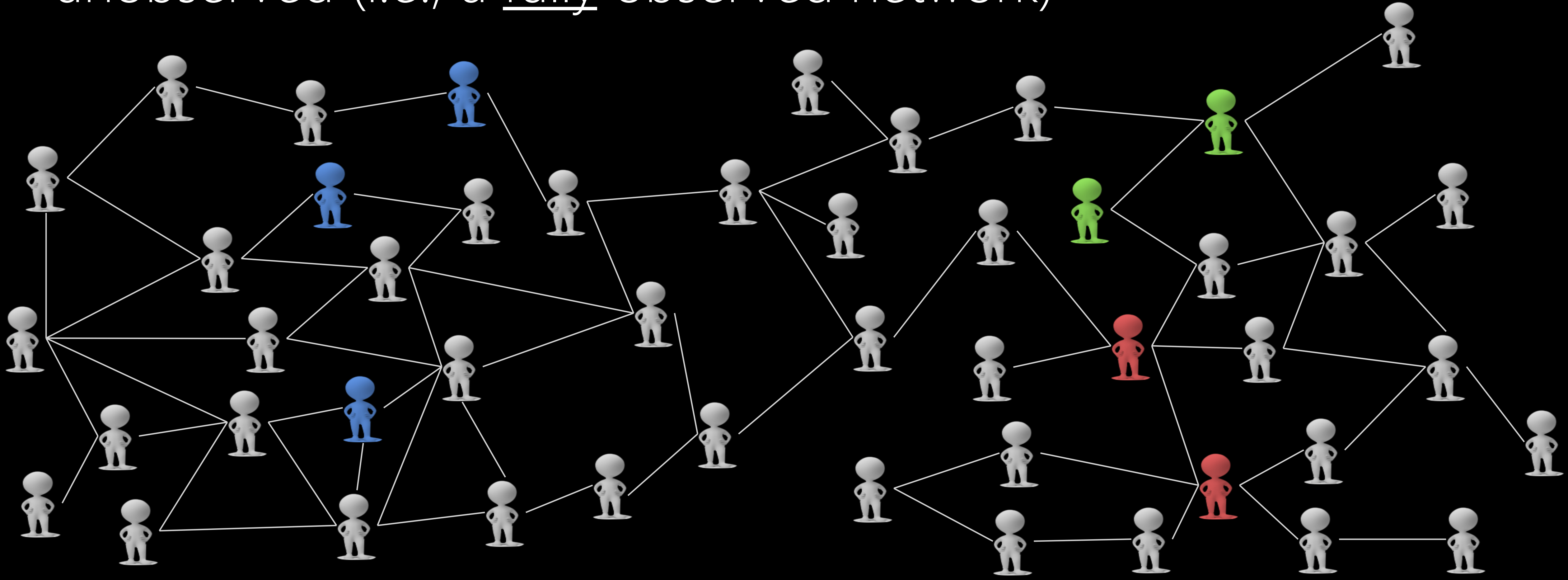
Comparison

Pattern #1 & #2: Collective Classification & Link Prediction

Nugget: active surveying – acquire label *and* neighbors

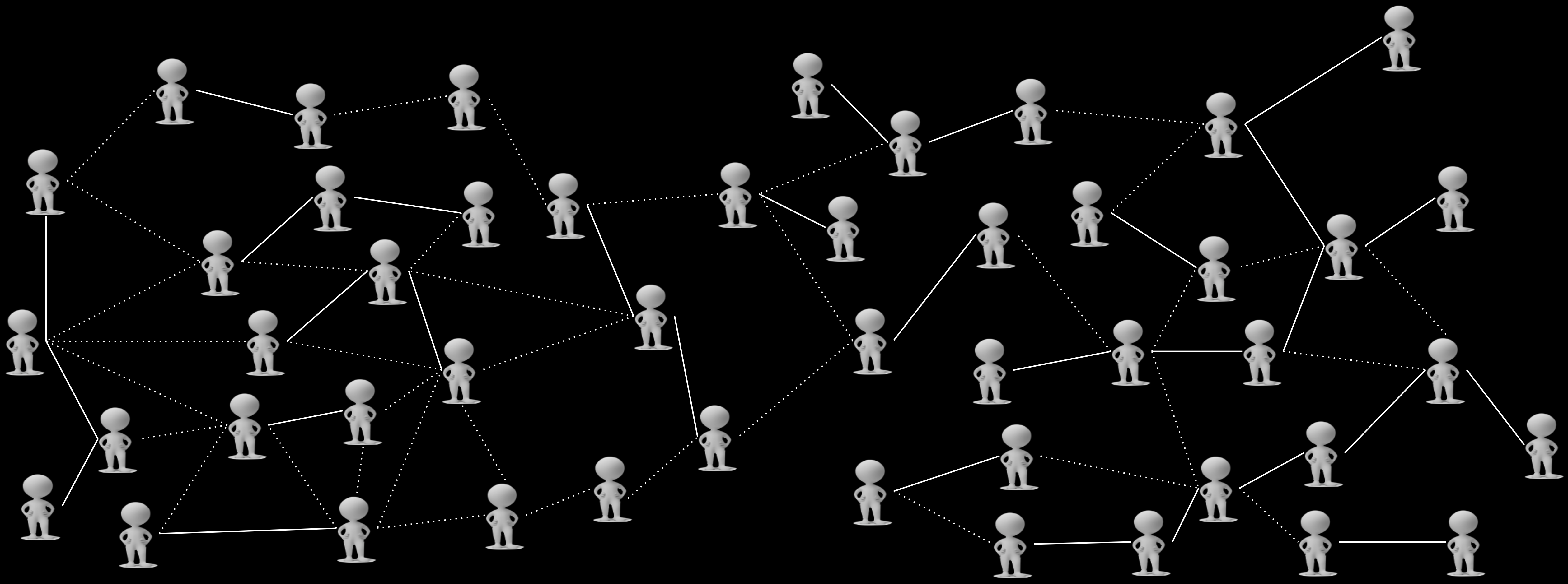
Sharara & Getoor IJCAI 2011; Namata et al., MLG 2012

Most previous work assumes that only the labels are unobserved (i.e., a fully observed network)



Label: 🟩 **Positive** 🟦 **Neutral** 🟥 **Negative**

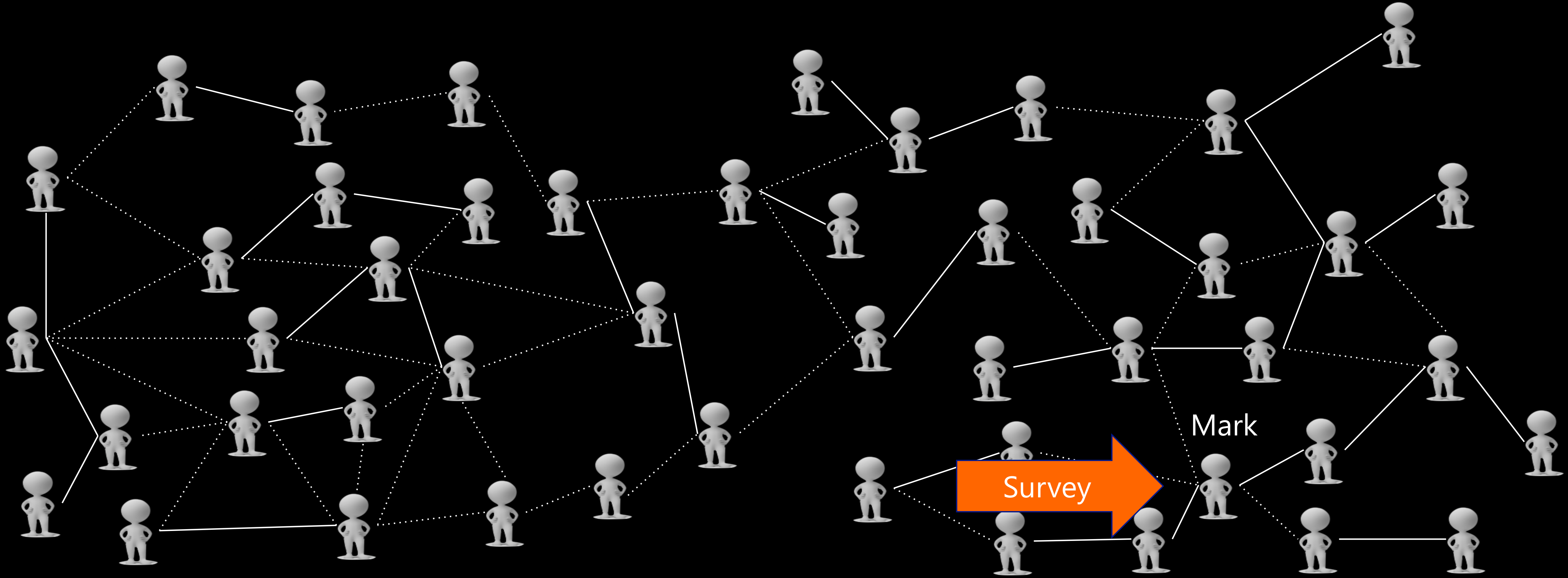
Network structure also often only partially observed



Label: 🟩 **Positive** 🟦 **Neutral** 🟥 **Negative**

Survey: Acquire the label and ego-network of a node

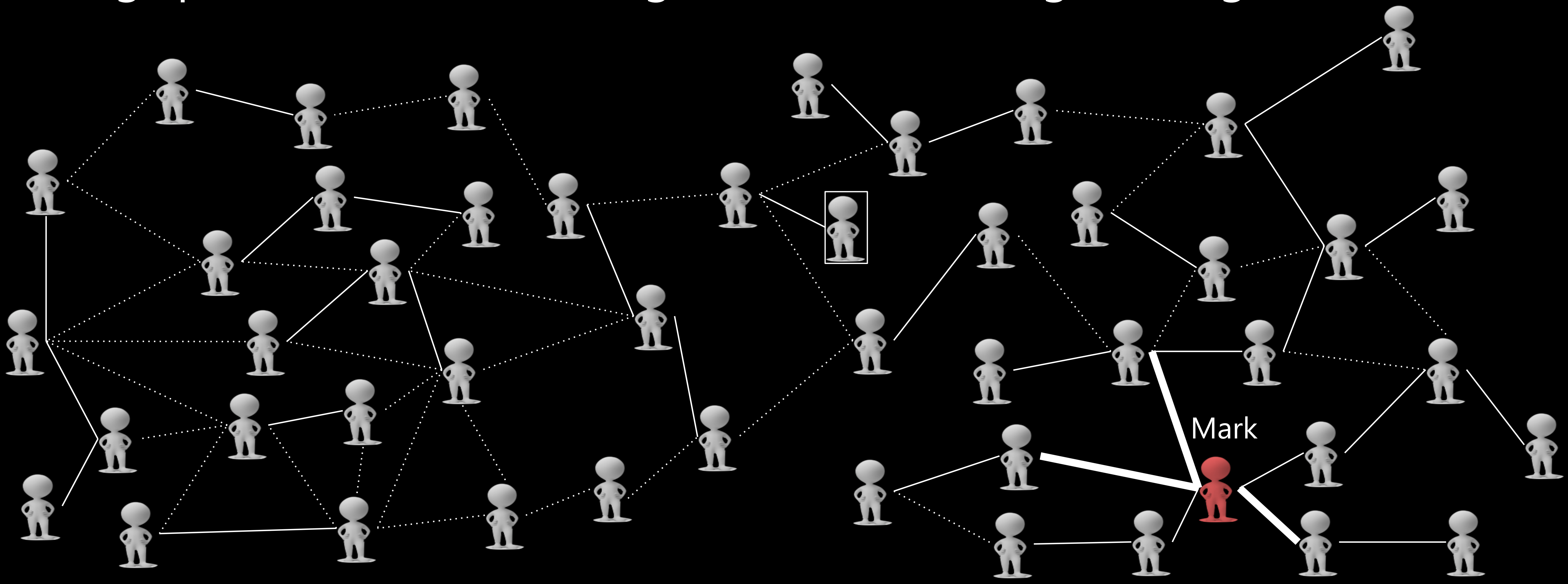
e.g., personal interview, targeted information gathering



Label: 🟩 **Positive** 🟦 **Neutral** 🟥 **Negative**

Survey: Acquire the label and ego-network of a node

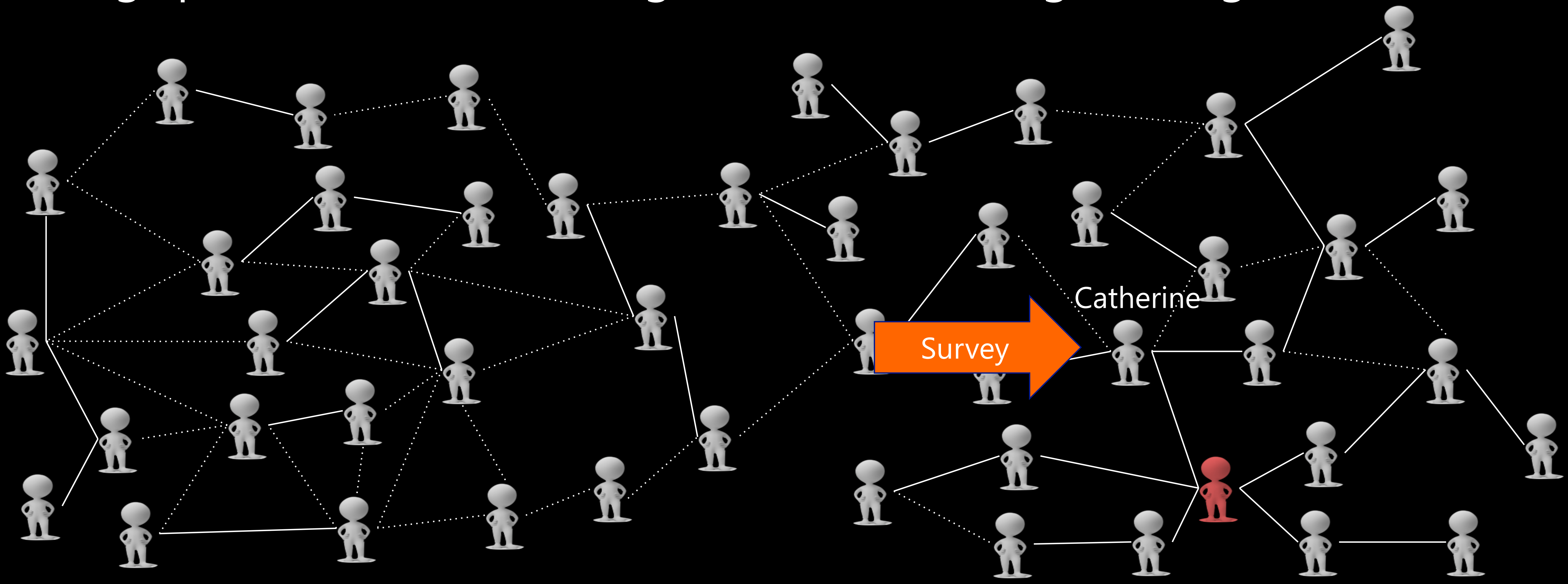
e.g., personal interview, targeted information gathering



Label: 🟩 **Positive** 🟦 **Neutral** 🟥 **Negative**

Survey: Acquire the label and ego-network of a node

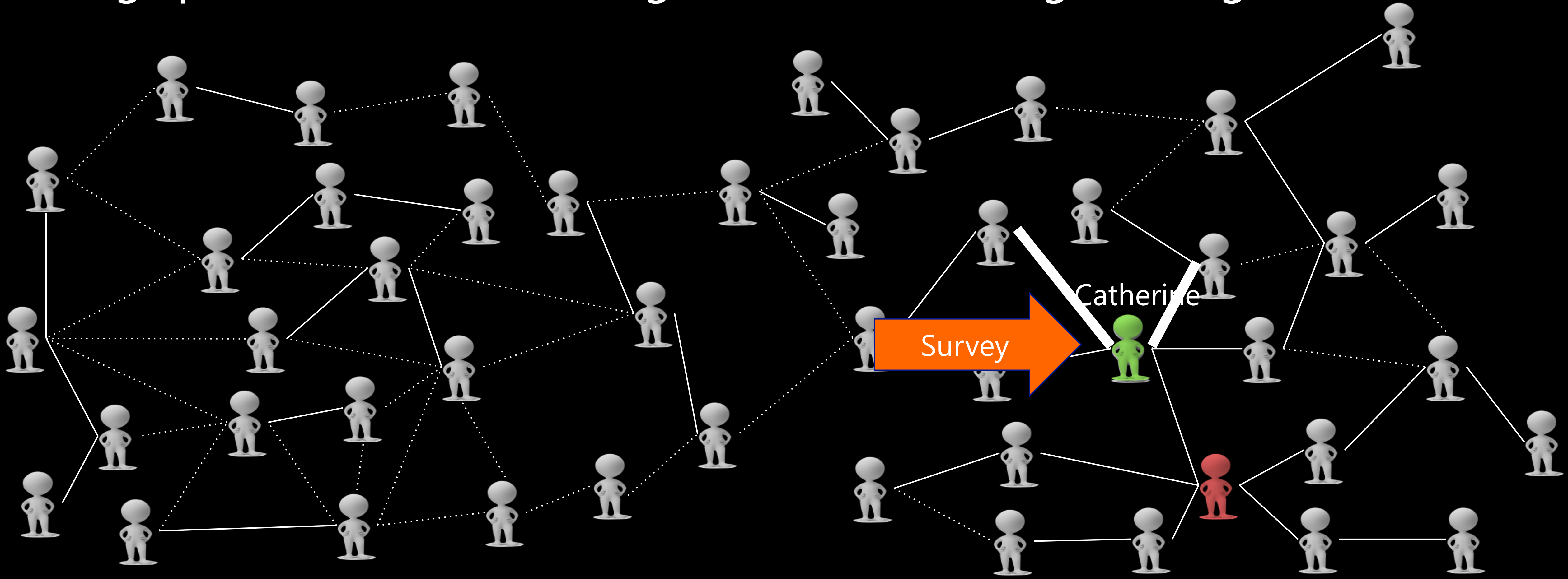
e.g., personal interview, targeted information gathering



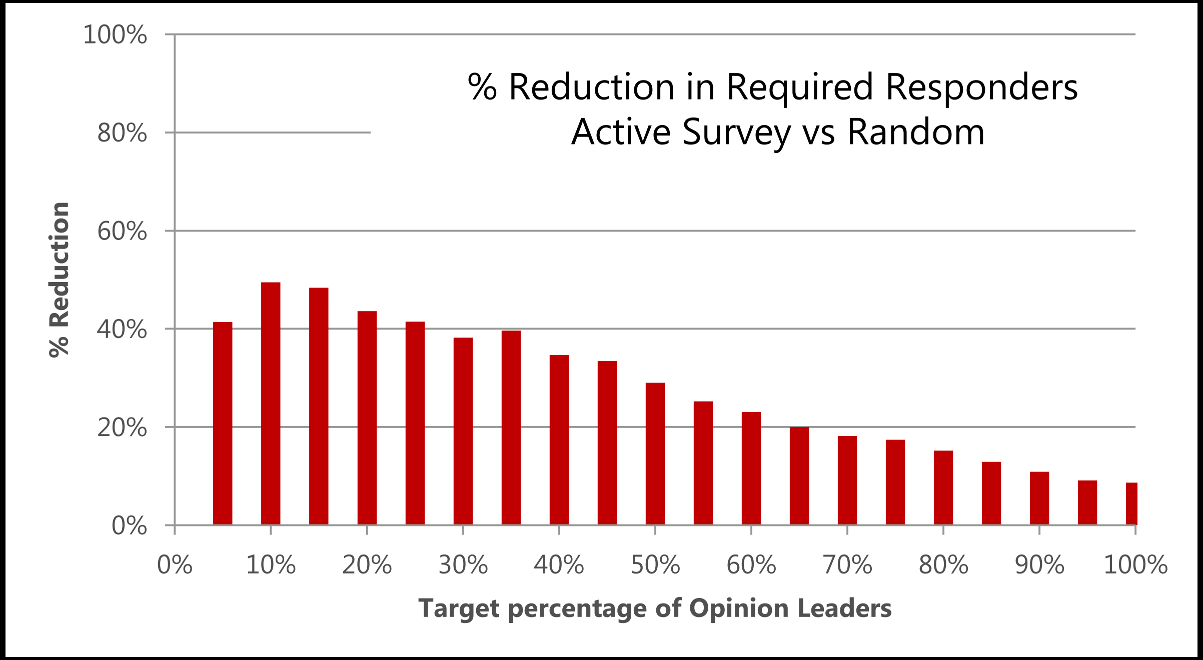
Label: 🟩 Positive 🟦 Neutral 🟥 Negative

Survey: Acquire the label and ego-network of a node

e.g., personal interview, targeted information gathering



Label: 🟩 **Positive** 🟦 **Neutral** 🟥 **Negative**


% Reduction in Required Responders
Active Survey vs Random
% Reduction
100%
80%
60%
40%
20%
0%
0%
10%
20%
30%
40%
50%
60%
70%
80%
90%
100%
Target percentage of Opinion Leaders

What’s different about graphs?

Unit of Interaction

Context

Comparison

Context

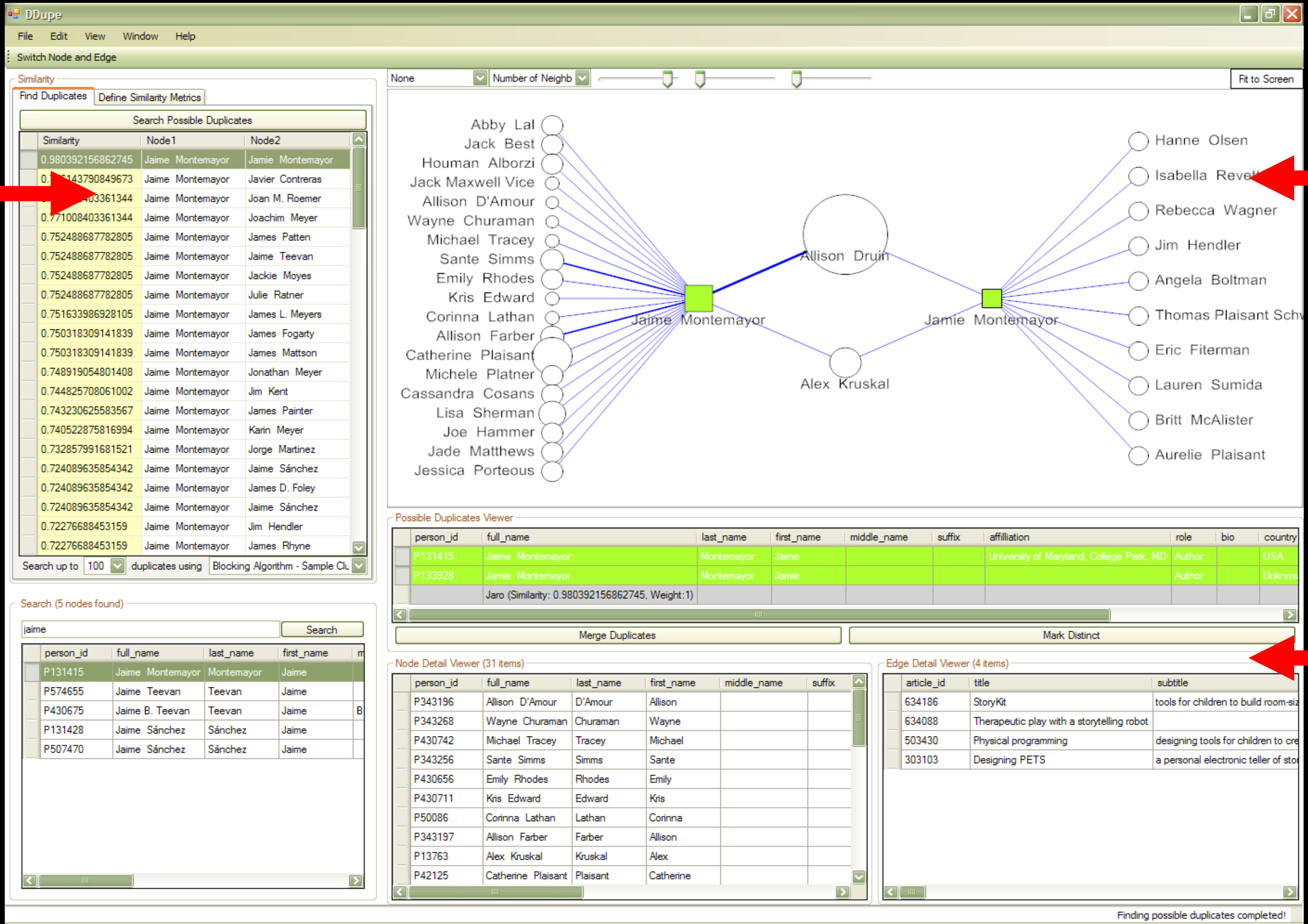
too little: single node

too much: whole graph

just right: relational context

D-Dupe: Interactive Entity Resolution Tool

Potential Duplicate Viewer



Similarity	Node1	Node2
0.980392156862745	Jaime Montemayor	Jamie Montemayor
0.[illegible]143790849673	Jaime Montemayor	Javier Contreras
[illegible]03361344	Jaime Montemayor	Joan M. Roemer
0.771008403361344	Jaime Montemayor	Joachim Meyer
0.752488687782805	Jaime Montemayor	James Patten
0.752488687782805	Jaime Montemayor	Jaime Teevan
0.752488687782805	Jaime Montemayor	Jackie Moyes
0.752488687782805	Jaime Montemayor	Julie Ratner
0.751633986928105	Jaime Montemayor	James L. Meyers
0.750318309141839	Jaime Montemayor	James Fogarty
0.750318309141839	Jaime Montemayor	James Mattson
0.748919054801408	Jaime Montemayor	Jonathan Meyer
0.744825708061002	Jaime Montemayor	Jim Kent
0.743230625583567	Jaime Montemayor	James Painter
0.740522875816994	Jaime Montemayor	Karin Meyer
0.732857991681521	Jaime Montemayor	Jorge Martinez
0.724089635854342	Jaime Montemayor	Jaime Sánchez
0.724089635854342	Jaime Montemayor	James D. Foley
0.724089635854342	Jaime Montemayor	Jaime Sánchez
0.72276688453159	Jaime Montemayor	Jim Hendler
0.72276688453159	Jaime Montemayor	James Rhyne

Search up to 100 duplicates using Blocking Algorithm - Sample Clu

Search (5 nodes found)

jaime

person_id	full_name	last_name	first_name	m
P131415	Jaime Montemayor	Montemayor	Jaime	
P574655	Jaime Teevan	Teevan	Jaime	
P430675	Jaime B. Teevan	Teevan	Jaime	B
P131428	Jaime Sánchez	Sánchez	Jaime	
P507470	Jaime Sánchez	Sánchez	Jaime	

Possible Duplicates Viewer

person_id	full_name	last_name	first_name	middle_name	suffix	affilliation	role	bio	country
P131415	Jaime Montemayor	Montemayor	Jaime			University of Maryland, College Park, MD	Author		USA
P133928	Jamie Montemayor	Montemayor	Jamie				Author		Unknow
	Jaro (Similarity: 0.980392156862745, Weight:1)								

Node Detail Viewer (31 items)

person_id	full_name	last_name	first_name	middle_name	suffix
P343196	Allison D'Amour	D'Amour	Allison		
P343268	Wayne Churaman	Churaman	Wayne		
P430742	Michael Tracey	Tracey	Michael		
P343256	Sante Simms	Simms	Sante		
P430656	Emily Rhodes	Rhodes	Emily		
P430711	Kris Edward	Edward	Kris		
P50086	Corinna Lathan	Lathan	Corinna		
P343197	Allison Farber	Farber	Allison		
P13763	Alex Kruskal	Kruskal	Alex		
P42125	Catherine Plaisant	Plaisant	Catherine		

Edge Detail Viewer (4 items)

article_id	title	subtitle
634186	StoryKit	tools for children to build room-siz
634088	Therapeutic play with a storytelling robot	
503430	Physical programming	designing tools for children to cre
303103	Designing PETS	a personal electronic teller of stor

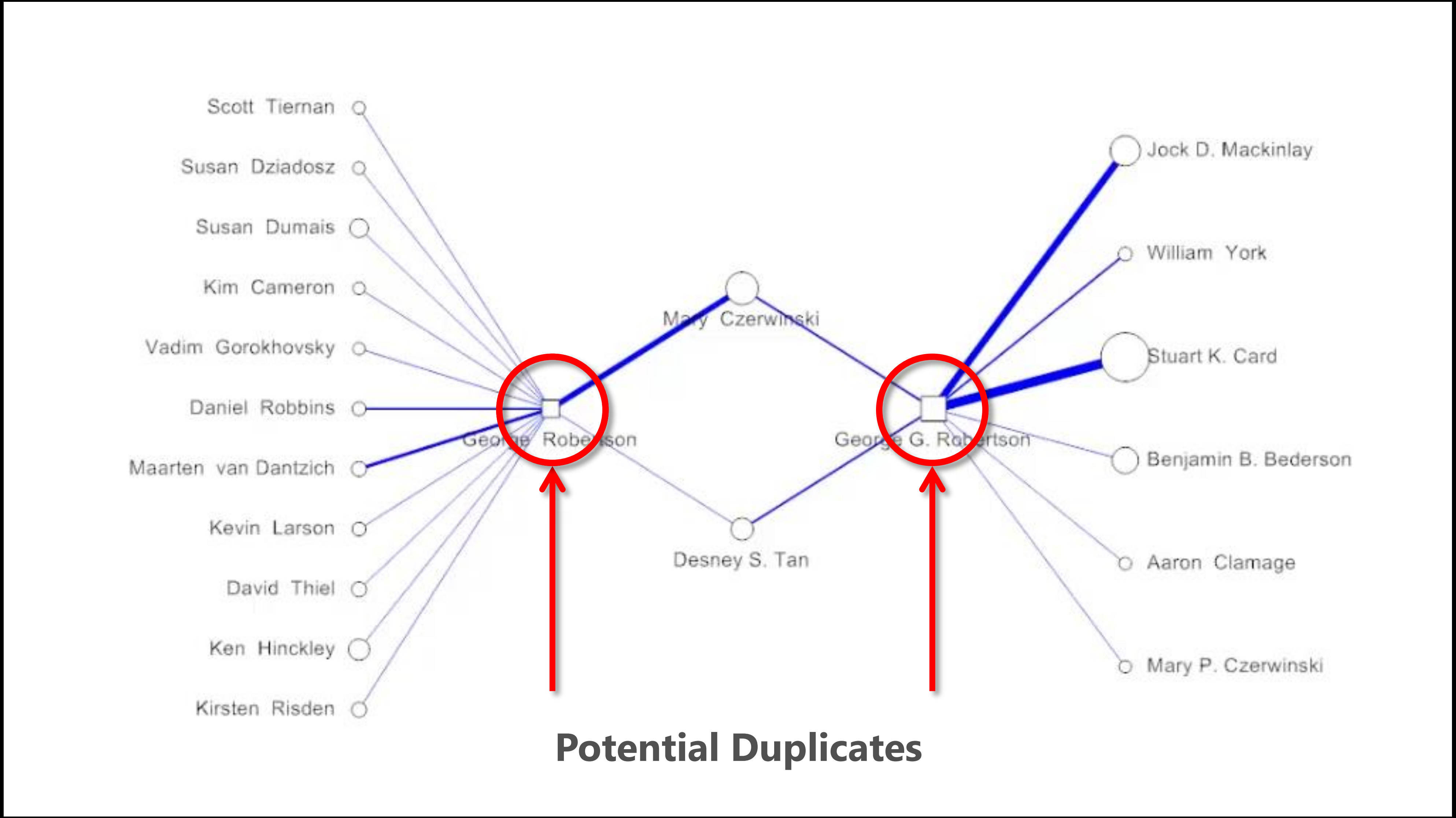
Finding possible duplicates completed!

Relational Context Viewer

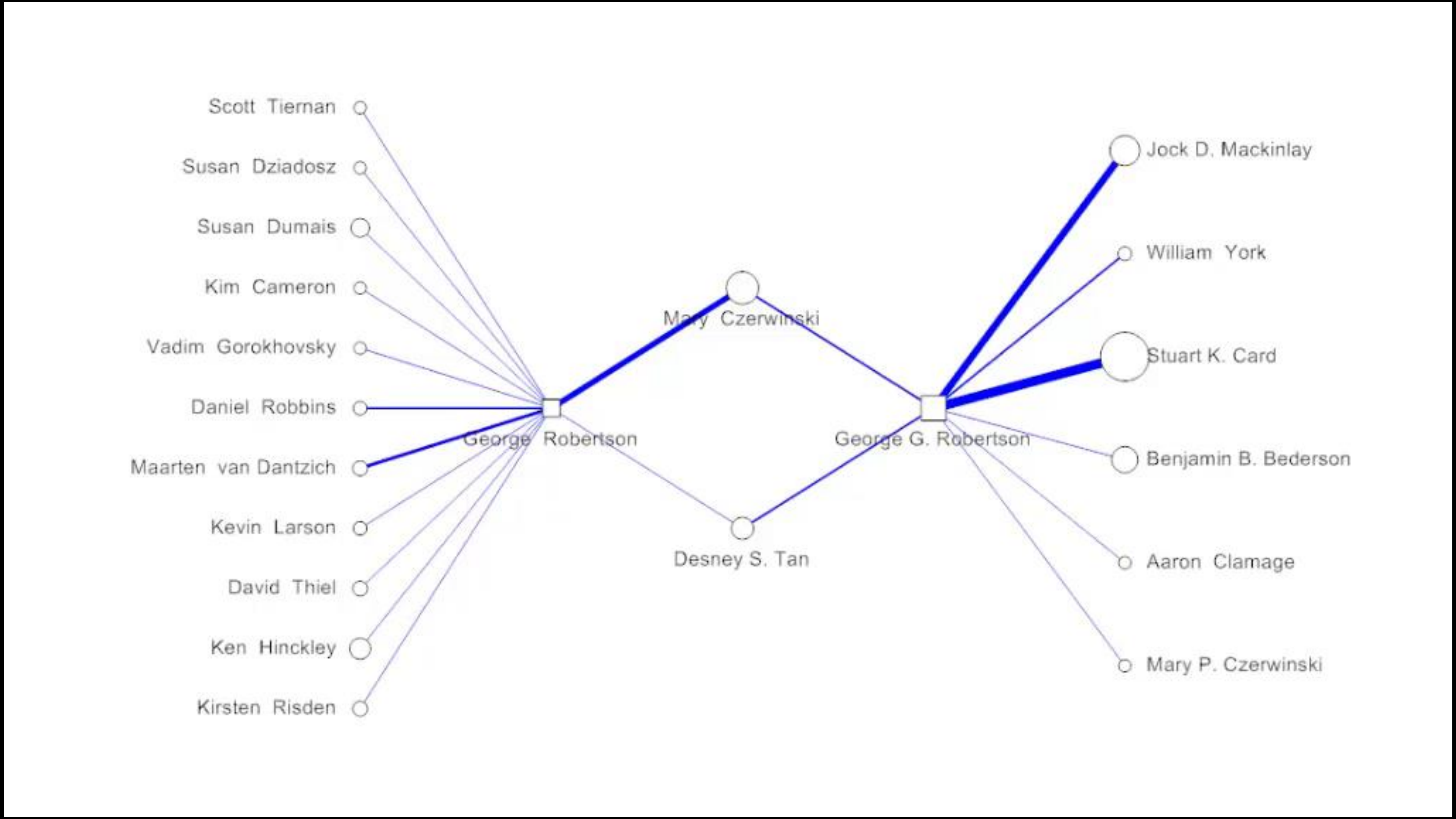
Data Detail Viewer

Kang, Getoor, Shneiderman, Bilgic, Licamele, TVCG 2008
http://www.cs.umd.edu/projects/linqs/ddupe

Nugget: Relational Context



Nugget: Relational Context



What’s different about graphs?

Unit of Interaction

Context

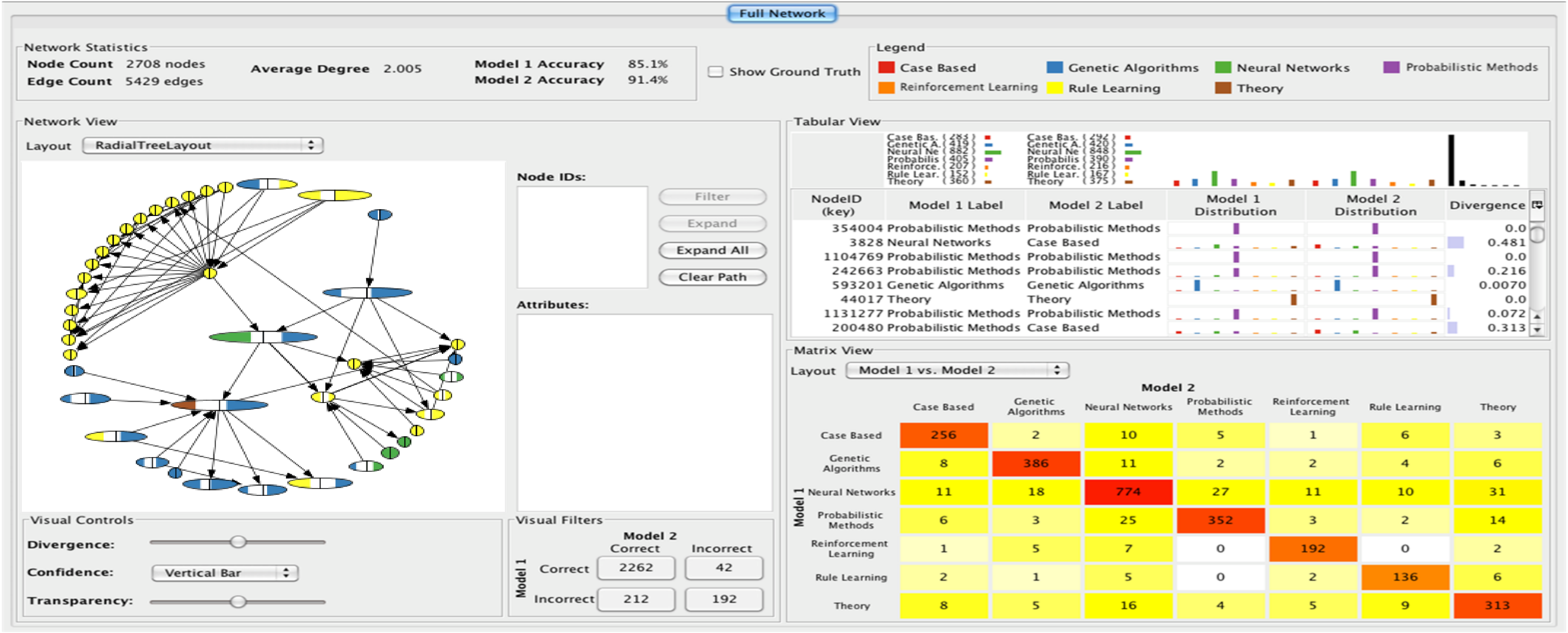
Comparison

Comparing ML Algorithms

Flat Data: confusion matrix

Graph Data: ?

G-Pare: Graph Comparison



Network Statistics

Node Count 2708 nodes

Edge Count 5429 edges

Average Degree 2.005

Model 1 Accuracy 85.1%

Model 2 Accuracy 91.4%

□ Show Ground Truth

Legend

🟥 Case Based · 🟦 Genetic Algorithms · 🟩 Neural Networks · 🟪 Probabilistic Methods · 🟧 Reinforcement Learning · 🟨 Rule Learning · 🟫 Theory

Network View

Layout: RadialTreeLayout

Node IDs:

Attributes:

Tabular View

Model 1		Model 2	
Case Bas.	(283)	Case Bas.	(292)
Genetic A.	(419)	Genetic A.	(420)
Neural Ne	(882)	Neural Ne	(848)
Probabilis	(405)	Probabilis	(390)
Reinforce.	(207)	Reinforce.	(216)
Rule Lear.	(152)	Rule Lear.	(167)
Theory	(360)	Theory	(375)

NodeID (key)	Model 1 Label	Model 2 Label	Model 1 Distribution	Model 2 Distribution	Divergence
354004	Probabilistic Methods	Probabilistic Methods			0.0
3828	Neural Networks	Case Based			0.481
1104769	Probabilistic Methods	Probabilistic Methods			0.0
242663	Probabilistic Methods	Probabilistic Methods			0.216
593201	Genetic Algorithms	Genetic Algorithms			0.0070
44017	Theory	Theory			0.0
1131277	Probabilistic Methods	Probabilistic Methods			0.072
200480	Probabilistic Methods	Case Based			0.313

Matrix View

Layout: Model 1 vs. Model 2

Model 1 \ Model 2	Case Based	Genetic Algorithms	Neural Networks	Probabilistic Methods	Reinforcement Learning	Rule Learning	Theory
Case Based	256	2	10	5	1	6	3
Genetic Algorithms	8	386	11	2	2	4	6
Neural Networks	11	18	774	27	11	10	31
Probabilistic Methods	6	3	25	352	3	2	14
Reinforcement Learning	1	5	7	0	192	0	2
Rule Learning	2	1	5	0	2	136	6
Theory	8	5	16	4	5	9	313

Visual Controls

Divergence:

Confidence: Vertical Bar

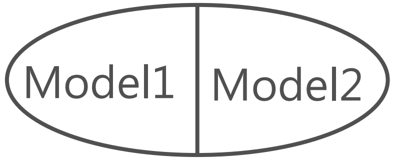
Transparency:

Visual Filters

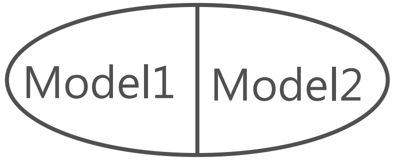
Model 1 \ Model 2	Correct	Incorrect
Correct	2262	42
Incorrect	212	192

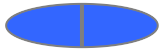
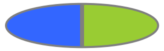
Sharara, Sopan, Namata, Getoor, VAST 2011
http://www.cs.umd.edu/projects/linqs/gpare

Nugget: Node Visualization

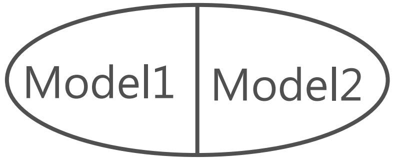


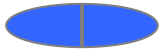
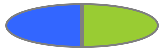
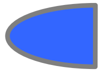
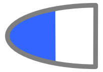
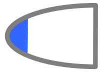
Nugget: Node Visualization



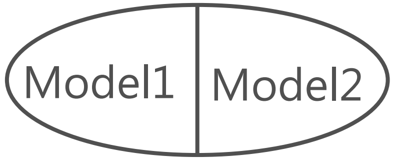
Color Coding	Predicted Label	🟦 Neutral 🟩 Positive  Agree  Disagree

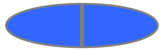
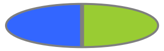
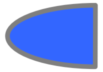
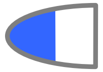
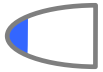
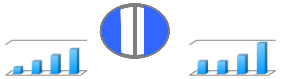
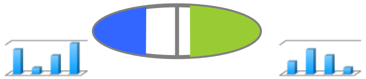
Nugget: Node Visualization



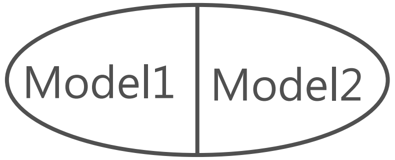
Color Coding	Predicted Label	🟦 Neutral 🟩 Positive  Agree  Disagree
Fill Area	Prediction Confidence	 High Confidence  Moderate Confidence  Low Confidence

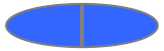
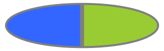
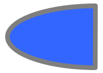
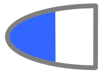
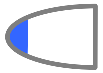
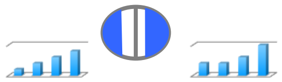
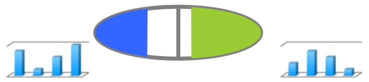
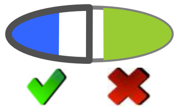
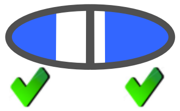
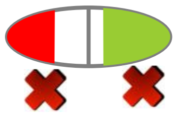
Nugget: Node Visualization



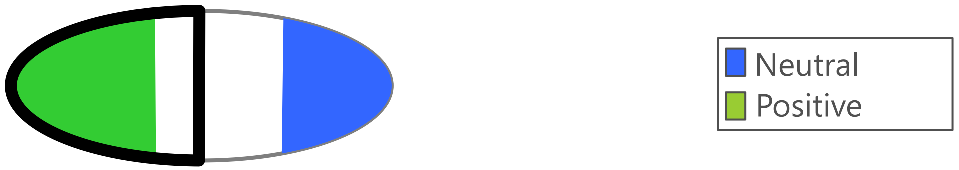
Color Coding	Predicted Label	🟦 Neutral 🟩 Positive  Agree  Disagree
Fill Area	Prediction Confidence	 High Confidence  Moderate Confidence  Low Confidence
Eccentricity	KL-Divergence	 

Nugget: Node Visualization



Color Coding	Predicted Label	🟦 Neutral 🟩 Positive  Agree  Disagree
Fill Area	Prediction Confidence	 High Confidence  Moderate Confidence  Low Confidence
Eccentricity	KL-Divergence	 
Border Highlighting	Ground Truth (Prediction Accuracy)	  

Nugget: Node Visualization



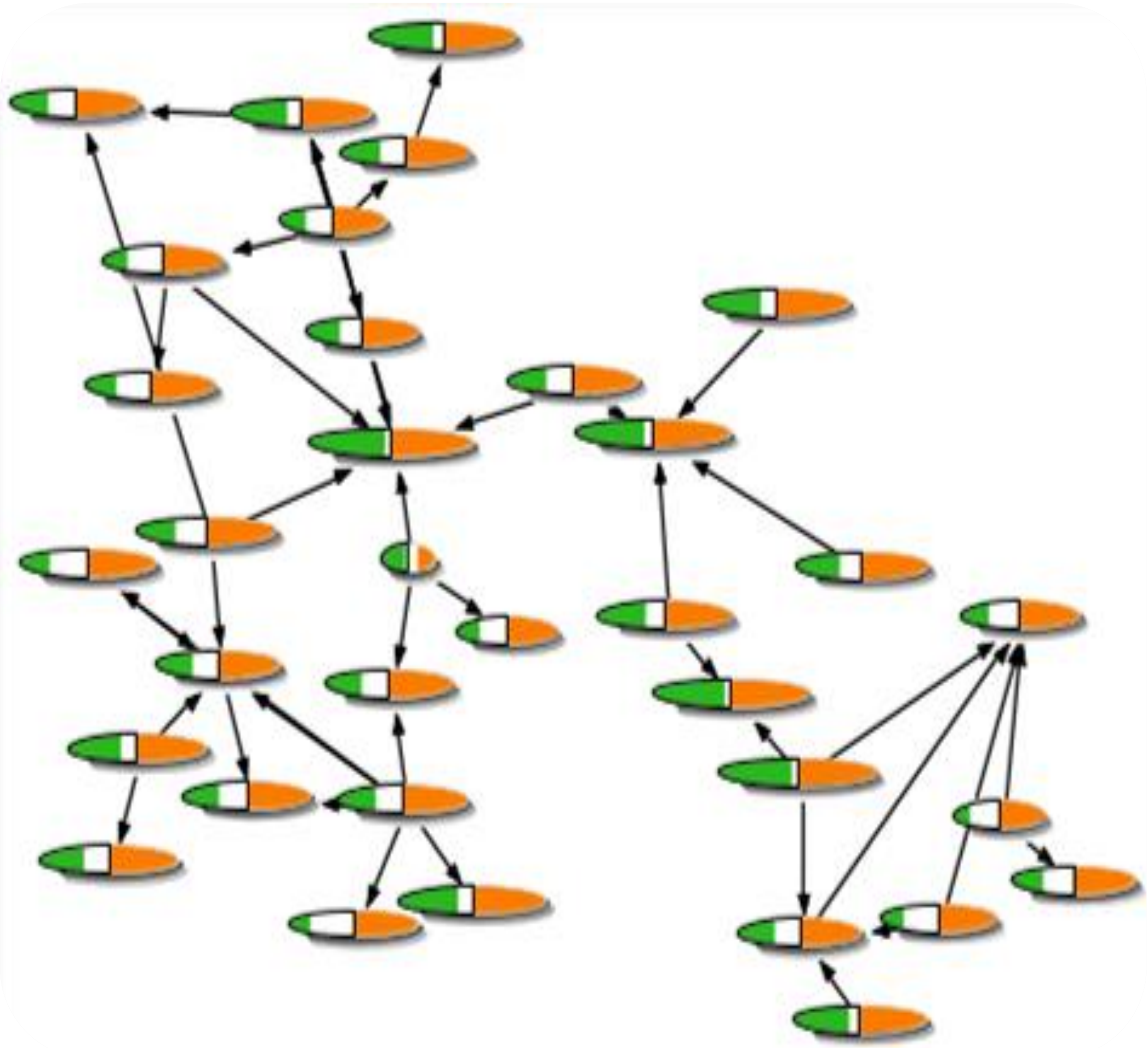
- Model 1 prediction: “Positive”
  Model 2 prediction: “Neutral”

- Model 1 is more confident in its prediction than Model 2

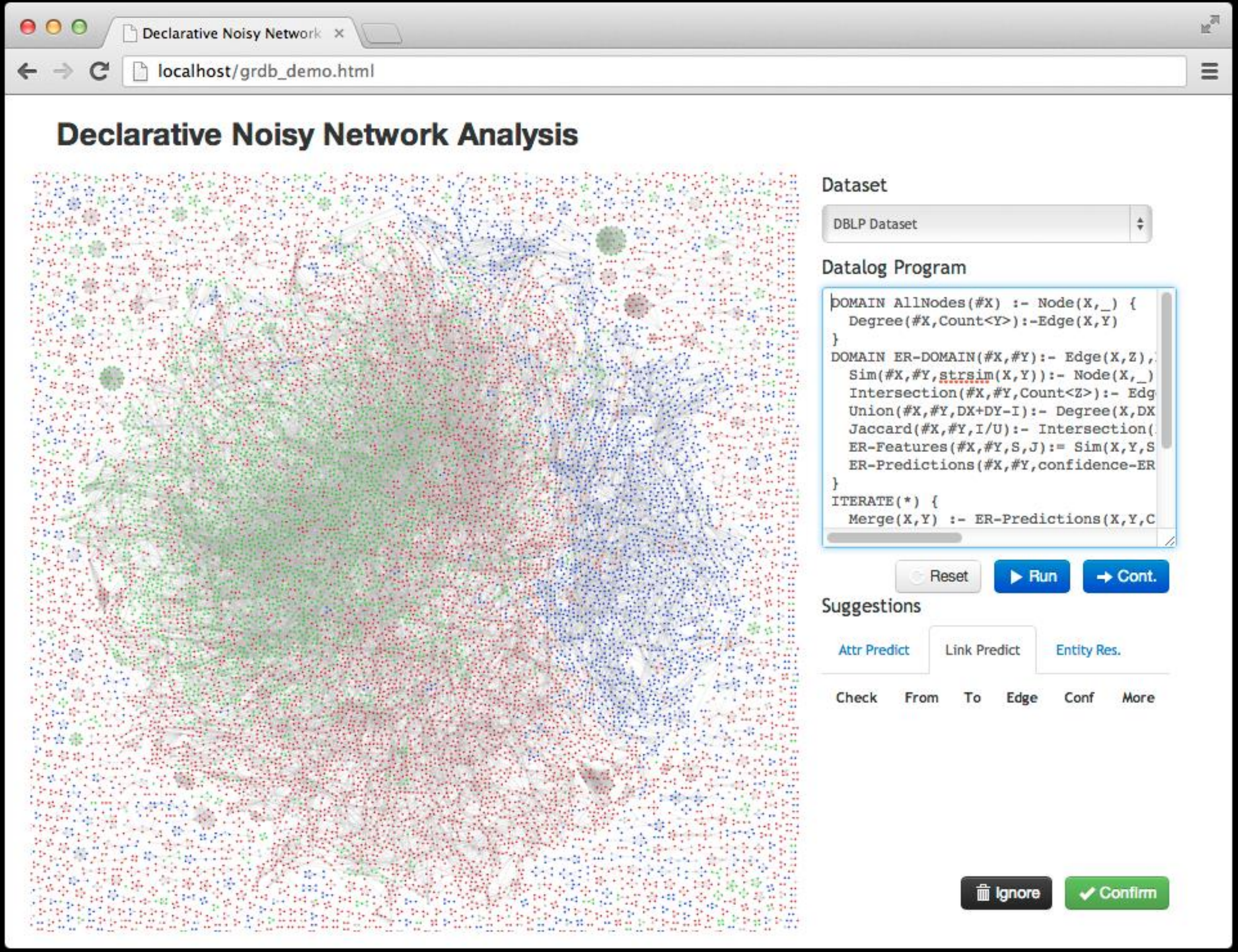
- Distributions of the two models vary significantly

- Model 1’s prediction matches the ground truth

Finding regions of disagreement



GrDB: Putting it all together, first steps…



Declarative Noisy Network Analysis

Dataset: DBLP Dataset

Datalog Program

```
DOMAIN AllNodes(#X) :- Node(X,_) {
  Degree(#X,Count<Y>):-Edge(X,Y)
}
DOMAIN ER-DOMAIN(#X,#Y):- Edge(X,Z),
  Sim(#X,#Y,strsim(X,Y)):- Node(X,_)
  Intersection(#X,#Y,Count<Z>):- Edg
  Union(#X,#Y,DX+DY-I):- Degree(X,DX
  Jaccard(#X,#Y,I/U):- Intersection(
  ER-Features(#X,#Y,S,J):= Sim(X,Y,S
  ER-Predictions(#X,#Y,confidence-ER
}
ITERATE(*) {
  Merge(X,Y) :- ER-Predictions(X,Y,C
```

Eldin Moustafa, Miao, Deshpande, Getoor, SIGMOD Demo 2013
http://www.cs.umd.edu/projects/linqs/grdb

Closing

State-of-the-Art: interaction unit, context and comparison important

Challenges: interaction/ML for complex tasks involving graphs is hard

Opportunities: creating common abstractions that work for both interaction for ML and ML for interaction