BIG Data is not flat
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?
What’s different about graphs?

Unit of Interaction
Context
Comparison
What’s different about graphs?

Unit of Interaction
Context
Comparison
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)
Network structure also often only partially observed
Survey: Acquire the label and ego-network of a node e.g., personal interview, targeted information gathering
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
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
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

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

Potential Duplicates
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

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

Pattern #1: Collective Classification
Nugget: Node Visualization

<table>
<thead>
<tr>
<th>Color Coding</th>
<th>Predicted Label</th>
<th>Ground Truth (Prediction Accuracy)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Model1</td>
<td>Model2</td>
</tr>
<tr>
<td>ECC</td>
<td>KL-Divergence</td>
<td>Border Highlighting</td>
</tr>
</tbody>
</table>
Nugget: Node Visualization

<table>
<thead>
<tr>
<th>Color Coding</th>
<th>Predicted Label</th>
<th>Prediction Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model1</td>
<td>Model2</td>
<td></td>
</tr>
</tbody>
</table>

- **Color Coding**: Predicted Label
  - **Model1**: Neutral
  - **Model2**: Positive
- **Fill Area**: Prediction Confidence
  - High Confidence
  - Moderate Confidence
  - Low Confidence

Pattern #1: Collective Classification
Nugget: Node Visualization

Pattern #1: Collective Classification

<table>
<thead>
<tr>
<th>Color Coding</th>
<th>Predicted Label</th>
<th>Model1</th>
<th>Model2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>![Neutral](Image</td>
<td><img src="Image" alt="Positive" /></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Fill Area</th>
<th>Prediction Confidence</th>
<th><img src="Image" alt="High Confidence" /></th>
<th><img src="Image" alt="Moderate Confidence" /></th>
<th><img src="Image" alt="Low Confidence" /></th>
</tr>
</thead>
<tbody>
<tr>
<td>Eccentricity</td>
<td>KL-Divergence</td>
<td><img src="Image" alt="Histogram" /></td>
<td><img src="Image" alt="Histogram" /></td>
<td><img src="Image" alt="Histogram" /></td>
</tr>
</tbody>
</table>
## Nugget: Node Visualization

<table>
<thead>
<tr>
<th>Color Coding</th>
<th>Predicted Label</th>
<th><img src="#" alt="Neutral" /> Positive</th>
<th><img src="#" alt="Agree" /></th>
<th><img src="#" alt="Disagree" /></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fill Area</strong></td>
<td>Prediction Confidence</td>
<td><img src="#" alt="High Confidence" /></td>
<td><img src="#" alt="Moderate Confidence" /></td>
<td><img src="#" alt="Low Confidence" /></td>
</tr>
<tr>
<td><strong>Eccentricity</strong></td>
<td>KL-Divergence</td>
<td><img src="#" alt="Pattern #" /></td>
<td><img src="#" alt="Collective Classification" /></td>
<td></td>
</tr>
<tr>
<td><strong>Border Highlighting</strong></td>
<td>Ground Truth (Prediction Accuracy)</td>
<td><img src="#" alt="Neutral" /></td>
<td><img src="#" alt="Positive" /></td>
<td><img src="#" alt="Agree" /></td>
</tr>
</tbody>
</table>
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...

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