Identifying Suspicious Bidders Utilizing Hierarchical Clustering and Decision Trees

Benjamin Ford
Master of Science Student
Computer and Information Science Department
University of Massachusetts Dartmouth
Email: BenjaminFordCIS@gmail.com

Acknowledgement

- Thesis Advisor
  - Haiping Xu, Associate Professor
- Collaborator
  - Iren Valova, Professor
- Sponsor
  - National Science Foundation

Number of eBay listings from 2008-2010

Online Auction Fraud

- Types of auction fraud
  - Non-delivery
  - False advertising
  - Bid shielding
  - Shill bidding

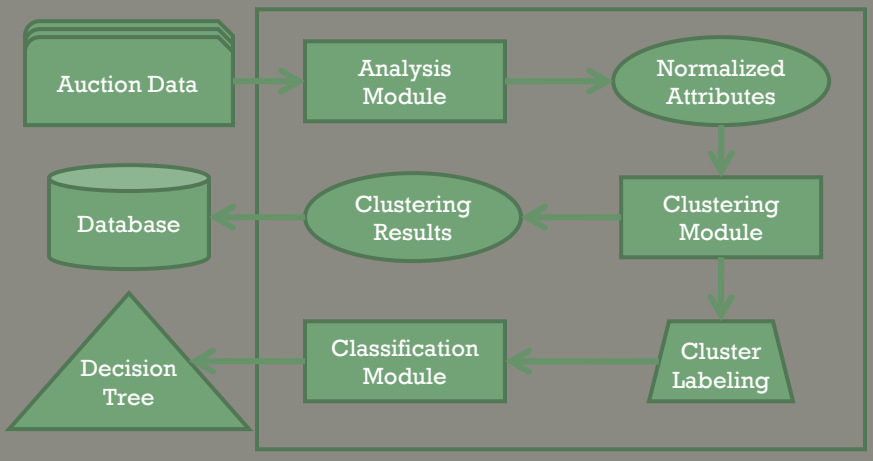
Shill Bidding

- Difficult to detect
  - Non-obvious, unlike other types of fraud
  - Collusion
  - Online anonymity
  - Multiple ways to participate in an auction
- Examples of warning signs
  - High bid amount(s) in beginning of auction
  - Bidding very close to beginning of auction
  - Bid unmasking

Detecting Suspicious Behavior

- Shill verification for every bidder infeasible
  - Numerous auctions for every bidder
- Most bidders are not shills
  - Faster to check for shill suspects and then verify
- Feature extraction to enable data mining

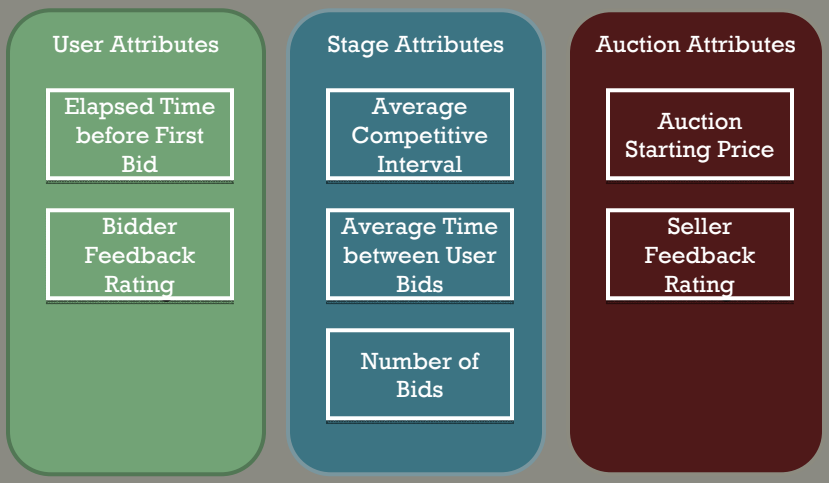
Identification Framework



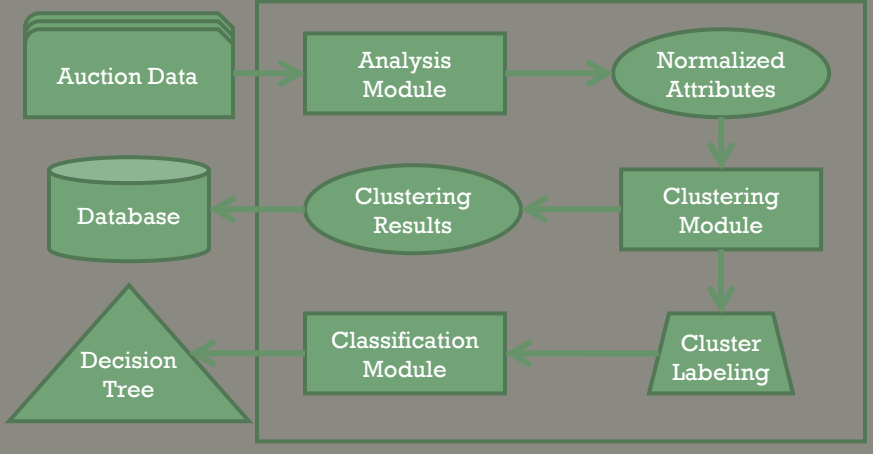
Measuring Bidder Behavior

- Series of measurable attributes
- Easily obtainable
  - Auction's bidding history
- Allows for more sophisticated techniques
  - Decision trees
  - Neural networks
  - Support vector machines

Examples of Bidder Attributes



Identification Framework



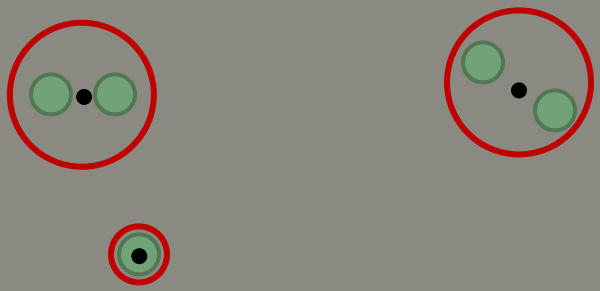
Hierarchical Clustering

- Group similar bidders based on behavior
- Similarity measure
  - Centroid clustering
  - Centroid = vector average of cluster's members
  - Similarity = dot product of 2 centroids

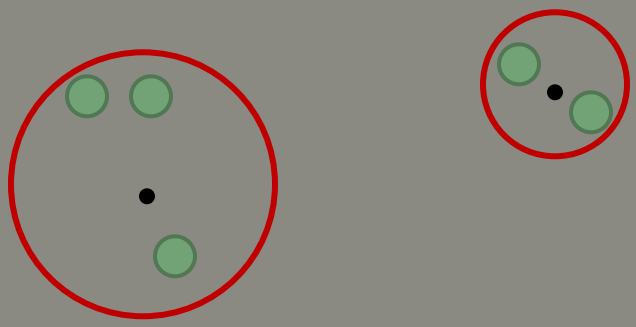
Cluster Generation Demo

Cluster Generation Demo

Cluster Generation Demo



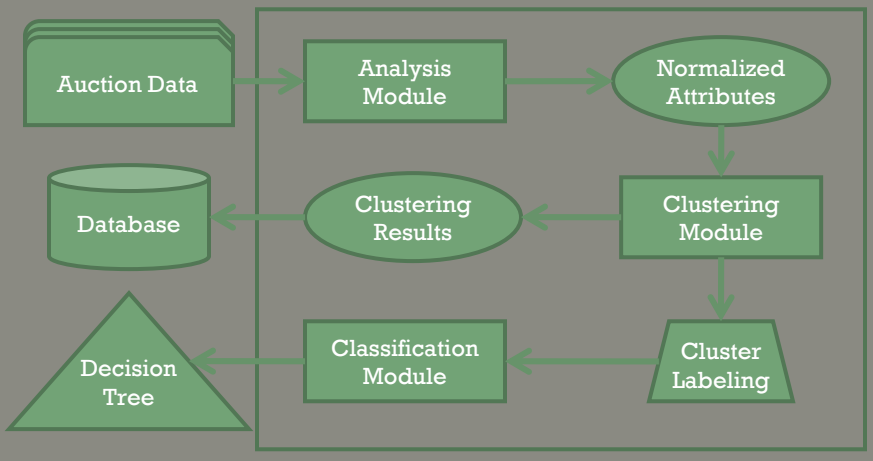
Cluster Generation Demo



Additional Considerations

- Minimum similarity cutoff
  - Terminates combination
- Normalization of attribute values
- Attribute weighting
  - Reflect relative importance of attributes

Identification Framework



Decision Tree

- Classification approach
- Information gain
  - Calculate the most significant attribute
- Gain ratio used to overcome overfitting

$$E(X) = \sum_{i=1}^{n} - P(v_i) * \log_2 P(v_i)$$

$$GAIN(X) = E(Y) - \sum_{i=1}^{n} P(X = v_i) * E(Y \mid X = v_i)$$

$$GAINRATIO(X) = \frac{GAIN(X)}{E(X)}$$

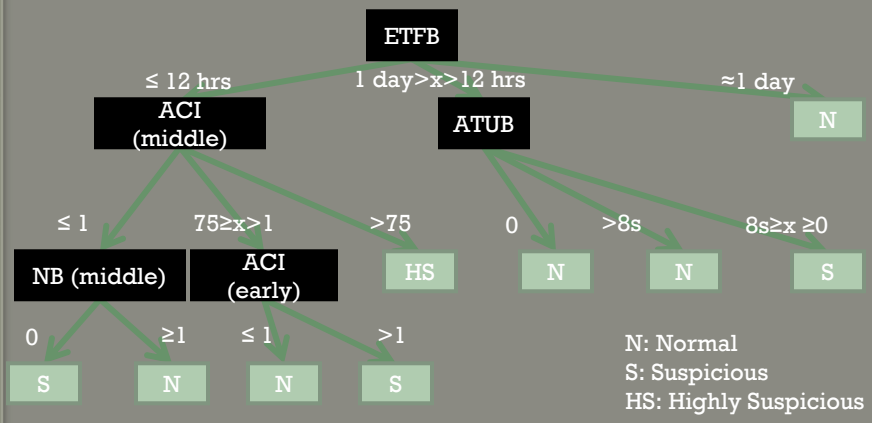
Case Study

- Real eBay data: "Used Playstation 3"
  - Collected over period of 30 days
  - Two separate datasets: 1 and 7 day durations
- 3-fold cross validation process
  - Fold 1 = First 1/3 test set
- Clustering minimum similarity cut-off = 86.9%

Labeling Results: 1-Day Fold 1

Cluster	Size	Class	Description
1	63%	Normal	Bids placed very late in auction (later middle stage or final stage).
2	<1%	Highly Suspicious	Very high bidding amounts in middle stage.
3	4%	Suspicious	Bids placed close together in middle stage. Possible bid unmasking.
4	9%	Normal	Few bids placed in the middle stage of auction.
5	8%	Normal	Similar to cluster 4, but bids placed later in the middle stage.
6	1%	Suspicious	Bids placed fairly early in auction.
7	<1%	Normal	Few bids placed in the middle stage of auction.
8	<1%	Highly Suspicious	Highest bid amounts in the middle stage.
9	1%	Suspicious	Bids placed close together in the middle stage. Possible bid unmasking.
10	<1%	Suspicious	Bids placed fairly early and bids placed close together in middle stage.
11	<1%	Highly Suspicious	Bids placed in quickest succession in the middle stage. Possible bid unmasking.
12	11%	Suspicious	Bids placed very early in auction (early stage).
13	<1%	Suspicious	Moderate number of bids in early stage.
14	<1%	Suspicious	Bids placed close together in early stage. Possible bid unmasking.
15	<1%	Highly Suspicious	Bids placed in quickest succession in the early stage. Possible bid unmasking.
16	<1%	Highly Suspicious	Highest number of bids in the early stage.

Decision Tree Result 1-Day Fold 1



Decision Tree Performance

1-Day Dataset, Fold 1, Decision Tree Statistics		
Training Set	Testing Set Size	Test Results
81% Normal 18% Suspicious 1% Highly Suspicious	614 data points	94% Correct 6% Incorrect

Conclusions & Future Work

- Quantified bidder behavior
- Grouped bidders based on behavior
- Created a decision tree to efficiently identify shill suspects
- Future work
  - More precise classifiers
    - Neural networks
    - Support vector machines
  - Stage-based classifiers

Questions?

For more information, refer to our project homepage at:

http://www.cis.umassd.edu/~hxu/Projects/ATM/