Analysis of Bidding Networks in eBay: Aggregate Preference Identification through Community Detection

R. Kang-Xing Jin[†]
Facebook, Inc.
Palo Alto, CA 94301 USA
rjin@post.harvard.edu

David C. Parkes and **Patrick J. Wolfe**
School of Engineering and Applied Sciences
Department of Statistics, Harvard University
Cambridge, MA 02138 USA
{parkes, patrick}@eecs.harvard.edu

Abstract

Statistical analysis of networks plays a critical role in the context of economics and the social sciences. Here we construct a bidding network to represent the behavior of users of the eBay marketplace. We study the eBay markets for digital cameras and liquid crystal display screens, and employ network analysis to identify aggregate structure in bidder preferences. The network that we construct associates auctions with nodes, and weighted edges between nodes capture the number of bidders competing in a pair of auctions, where said bidders ultimately win in only a single auction. We show that current community detection methods applied to this network allow for the identification of goods that are considered substitutes and complements, and thus the identification of aggregate preference information. In closing we suggest additional opportunities as well as challenges for the analysis of structured data in electronic markets.

Introduction

As technology enables the collection of ever more vast and diverse data sets, methods of unsupervised learning continue to grow in importance. Nowhere is this more readily apparent than in modern-day e-commerce networks, where the burgeoning study of behavior and preferences continues to gain in importance (Bapna 2004; Reichardt & Bornholdt 2005; Shah *et al.* 2003; Yang *et al.* 2003). However, the mathematical and statistical models and methodologies to support these studies lag significantly behind at present; existing algorithms do not allow researchers to ask and answer the new scientific questions engendered by these modern data types.

In particular, such data sets typically exhibit relational and temporal structure spanning a variety of scales. Here we consider a data set derived from the eBay online auction site. We collect information about the bidder activity in two markets, namely the market for branded digital cameras and the market for liquid crystal display (LCD) screens. From this information we extract relational data by looking for bidders who participate in multiple auctions. This provides a network representation of the market, in which auctions are associated with nodes, and weighted edges between nodes capture the number of bidders competing in a pair of auctions, where said bidders ultimately win in only a single auction.

Given such a "bidding network," we adopt community detection algorithms to identify aggregate preference information. Our main idea is that behavioral information about bidders can provide, in the aggregate, information about which auctions' goods are viewed as substitutes and which auctions' goods are viewed as complements. We demonstrate that community detection methods applied to this network can identify substitutes and complements goods, and thus aggregate preference information about the bidder population. We close with a discussion about the need for principled, Bayesian-inference based methods to study the structure of behavioral data in economic contexts.

The Problem: Aggregate Preference Identification

A central problem in the analysis of economic systems is to identify the preferences of the actors within the economy. Armed with information about preferences, one can then study both the allocative efficiency of market institutions, as well as propose new institutions in order to improve market efficiency. We are interested in methods that can be leveraged to provide robust information about *common structure* in user preferences.

Specifically, on eBay we are interested in the following two questions:

- Can *substitutes goods* be automatically identified from bidder behavior?
- Can *complements goods* be automatically identified from bidder behavior?

Roughly, two items are *substitutes* if a typical bidder with value for one item also has value for the other item.[1] Examples of substitutes goods include two models of 19" LCD monitors. Two items are *complements* if a typical bidder has superadditive value for the pair of items (and the demand for

[†]This work comprised the first author's senior thesis work while a student at Harvard University (Jin 2006).

[1]A more technical definition can be provided, whereby preferences satisfy gross substitutes if the demand for one of the items does not decrease when the price of the other item increases (Lehmann, Lehmann, & Nisan 2001).

one item might fall if the price on the other item increases.) Examples of complements goods include a digital camera and a memory stick, or a flight and matching hotel rooms.

This macro-level, structural information on the preferences of participants in a marketplace can be useful for a number of reasons. For instance:

- Knowledge of goods that are natural substitutes, or complements, can influence the design of user interfaces;
- Manufacturers can learn from revealed preference information about the products that are viewed as substitutes in the marketplace;
- In complex markets, it can be useful to pre-bundle complements (and sometimes substitutes; e.g., in the case of information goods) to improve revenue and/or efficiency properties and simplify elicitation;
- Such knowledge can also be used to enhance cross-selling (in the case of complements); e.g., attempting to sell digital memory in addition to a camera.

The question of substitutes is also in part motivated by the related problem of *categorization*: eBay has millions of widely varying items—for example, antiques, cars, real estate, and electronics, among others—so what is a scalable way to organize these items into categories? Categorization has been shown to have a significant positive effect on monthly sales (Lohse & Spiller 1998), and an important part of the success of eBay (Weiss, Capozzi, & Prusak 2004). Indeed, items with precise categories tend to attract more bidder traffic (Hahn 2001), and allows for search via category-specific parameters (McGuinness 2001), such as size, model and brand.

Related Work

In application domains, the analysis of social networks is well established (Wasserman & Faust 1994); recent work has begun to focus on model-based clustering for these networks (Handcock, Raftery, & Tantrum 2007). A number of other studies also treat classification and clustering in relational data (Taskar, Segal, & Koller 2001; Kemp, Griffiths, & Tenenbaum 2004). However, to the best of our knowledge, the application of such techniques to the economic analysis of electronic markets is a very recent development.

Community detection in networks is a growing area of research because it can help to reveal underlying structure (Palla *et al.* 2005). For example, Flake et al. (Flake *et al.* 2002) found that web pages tend to cluster into communities of semantically similar pages; community structure has also been examined in social and biological networks (Girvan & Newman 2002).

The empirical study of Internet auctions is a relatively new field, and most work has been done in the past five years. This fact is not too surprising, since online auction websites have only reached prominence recently. To our knowledge, no prior work has addressed the questions of identifying natural *substitutes* or natural *complements* amongst the goods in the eBay market.

The closest related work is a study of bidder communities on eBay (Reichardt & Bornholdt 2005), which we believe to be the first study of communities in a network generated from an online auction site.[2] The authors logged all data over a 12-day period on the German eBay.de website. They then generated a network with bidders as nodes and edges drawn between any two bidders who bid in the same auction. Next, they applied a community detection algorithm to the network. They found 7 "major" communities and noted that these communities tended to correspond to auctions in specific eBay-defined high-level goods categories. For example, one community consisted of bidders who primarily participated in the Toy Models and Toy categories. From these data, they concluded that bidders tend to limit their activities to general categories of goods. We ask different questions in our work (i.e., whether one can identify aggregate preference structure for specific goods within a category), and construct a different network to represent the eBay market.

From eBay to Bidding Networks

In this section we describe how we construct a bidding network from eBay data.

eBay Auction Data

We collect data by searching closed listings on eBay.com. We have developed harvesting scripts written in Perl to "scrape" eBay data, which is then stored in a Mysql database.[3] The information in the data set is multi-faceted and includes the following elements for each auction:

1. Title of auction, name of seller, type of auction, reserve price, reputation of seller.
2. Whether or not the item sold. The high bid in the auction, and the start and end time.
3. The name of each bidder, the time the bid was placed (to the bidder proxy) and the value of each bid.[4]

In addition, for the two eBay markets that we have studied we have collected descriptive information, including technology, brand information, etc. from an online source.

We collected data from two categories of goods. The first set (Canon) contains all auctions matching "Canon" in the Digital Cameras category over a period from Jan. 10, 2006 to Jan. 25, 2006. The second set (LCD) contains all auctions matching "LCD" in the Monitors and Projectors category over a period from Nov. 29, 2005 to Dec. 14, 2005. These markets were chosen because they are reasonably sized markets where there might be natural substitutes (specific models of cameras and specific sizes, brands, or models of LCDs).[5]

[2]Personal communication, M. E. J. Newman.

[3]These scripts were first developed in a classroom setting by co-author Parkes, and were subsequently improved by Aaron Roth.

[4]We do not have information on the value submitted by the *winning* bidder to her proxy, only a lower bound on this via the closing price.

[5]Note that these market choices define our set of possible substitutes for the purposes of this study, as we can only determine substitutes on which users bid.

The Canon set consisted of 6717 auctions, and the LCD set consisted of 11782 auctions. Similarly to Yang et al. (Yang *et al.* 2003), we found for both data sets that the distribution on the number of auctions in which a bidder participates appeared to follow a power law. In the Canon market, 8453 of the 12759 bidders (66%) participated in only one auction, and 2065 (16%) participated in only two. In the LCD market, 15650 of the 23801 bidders (66%) participated in only one auction, and 3883 (16%) participated in only two. A small number of bidders thus account for a disproportionate amount of bidding activity. The maximum numbers of auctions participated in by a single bidder were 61 and 171 for the Canon and LCD markets, respectively. The bidder who participated in 171 auctions in the LCD market averaged more than 12 per day!

Constructing a Bidding Network

Following data collection, we construct a graph to represent the semantic information about goods that is revealed through bidder behavior: auctions are nodes and an edge is drawn between an two auctions that share a common bidder. Thus, an edge conveys information about preferences: if one restricts edges to those that represent bidders that eventually win a single item then the presence of bidding across auctions provides revealed preference, in this case indicating that the items in the associated auctions are substitutes.

Preprocessing techniques are applied as follows in order to emphasize structure and remove noise due to extremal bidding behaviors: edges are weighted by the number of shared bidders in any two auctions; and edge weights due to bidders with bids less than some fraction f (e.g., 0.8) of the winning price in one or both of the auctions are removed. This proved important; for instance, the 171-auction bidder in the LCD data set, who represented less than 0.02% of the bidder population, was found to generate more than 10% of the graph edges. We do not want the community structure in our network to be dominated by any one bidder.

As an illustration of the resulting network, Figure 1 shows a spring-model energy-minimization representation of the largest maximally connected component of the Canon bidding network (for $f = 0.8$).

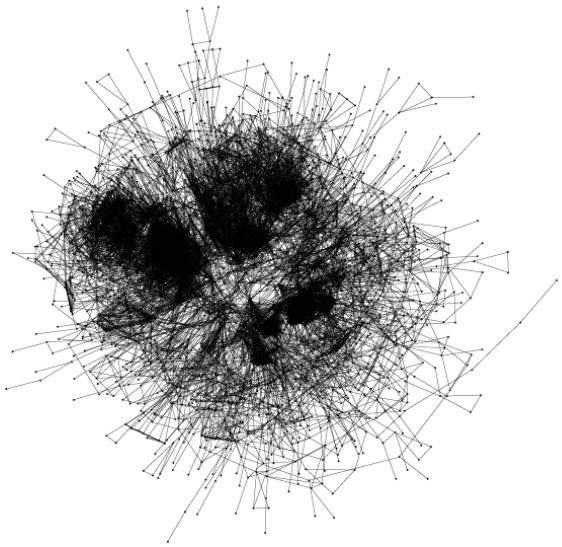

Figure 1: Largest maximally connected component of the Canon bidding network, derived from eBay data

Community Detection Methods

We apply a current, state-of-the-art community detection algorithm to this graph.

What is a Community?

Informally, a community is described as a subset of nodes that are connected more strongly to each other than to the rest of the network. For example, a social network of all university students in a locale, with edges defined by friendships, might naturally have communities defined along college lines.

To this end, one formal definition of a community is as follows (Reichardt & Bornholdt 2004): Given a graph G with N nodes and M edges, a community of n nodes and m edges is one satisfying $\frac{2m}{n(n-1)} > \frac{2M}{N(N-1)} > \frac{m_{nN}}{n(N-n)}$, where m_{nN} is the number of edges connecting the community to the rest of the network. Each of these terms represents a normalized edge density—the number of edges divided by the maximum number of edges possible (if the nodes were connected as a clique). The first inequality requires that within-community density be greater than the average network density, while the second inequality requires that the average network density be greater than the density of edges leaving the community. Loosely, a community should have a dense number of edges within the community and a sparse number of edges connecting it with the rest of the network.

Finding a Good Community Structure

For a given graph $G = (V, E)$, where $|V| = N$ and $|E| = M$, the community detection problem can be formalized as a partitioning problem subject to a constraint. Each $v \in V$ should be assigned to some partition c_i, for $i \in \{1, 2, \ldots, n_c\}$ where n_c is the number of communities. We want this partitioning to match our intuition as to what a good community is—that is, the partitions should satisfy some definition of community and/or maximize some metric that assesses a proposed community partitioning. The community detection problem is difficult because there may be multiple ways to divide any graph into acceptable communities, and furthermore, the number of "optimal" communities is often not known beforehand.

Modularity is a global metric that has been widely used to compare different community divisions and determine an "optimal" one (Newman 2004), and is defined as follows. Let e_{ii} be the fraction of all edges in the graph that lie *within* community i, and let $a_i = \frac{1}{2M} \sum_{v \in c_i} d_v$, where d_v is the degree of node v, denote the fraction of ends of edges in the

graph that are in community i. Then the modularity $Q(G, c)$ of a graph and given community division c is defined as $Q(G, c) = \sum_i (e_{ii} - {a_i}^2)$. This represents the fraction of edges in the network that fall within communities less the expected value of the fraction of edges in the network that fall within communities given the number of ends of edges in each community and if the edges had been assigned uniformly at random.

Adopting shorthand Q for this modularity metric, a value of $Q = 0$ indicates no community structure, while a value of Q approaching the maximum of 1 represents the presence of strong community structure.[6] Newman (Newman 2004) found that real world networks with strong community structure tend to have Q values of at least 0.3. The highest known value of Q in the literature for an unweighted real-world network is $Q = 0.75$, which came from a reviewer network from Amazon.com (Clauset, Newman, & Moore 2004).[7]

Having adopted this metric, we are now prepared to identify communities in a network. Given the largest maximally connected component of a network, we wish to find a community division that maximizes Q for this component. Since the number of possible partitions is exponential in the number of nodes, Newman (Newman 2004) propose a greedy forward-selection method ("Greedy Q") where one starts with each node in its own community and iteratively joins communities based on the greatest increase in Q. This method has been demonstrated to reliably identify communities in both artificial networks and real-world networks, and can be further optimized for computational efficiency (Clauset, Newman, & Moore 2004).

Figure 2 demonstrates the results of applying this algorithm to a subset of 10% the eBay LCD market, with nodes shown as auctions and edges drawn between any two auctions that share a common bidder.

Results: Substitutes Preferences

For each of our markets—Canon and LCD—we generated a network that used both price threshold edge filtering (at $f = 0.8$) and edge weighting, and applied our solution. We applied keyword extraction[8] to the communities and examined the keywords. We first discuss results for the Canon market, and then turn to the LCD market.

Canon Digital Cameras

Figure 3 illustrates the result of applying the methodology described above to our digital camera data set. The size

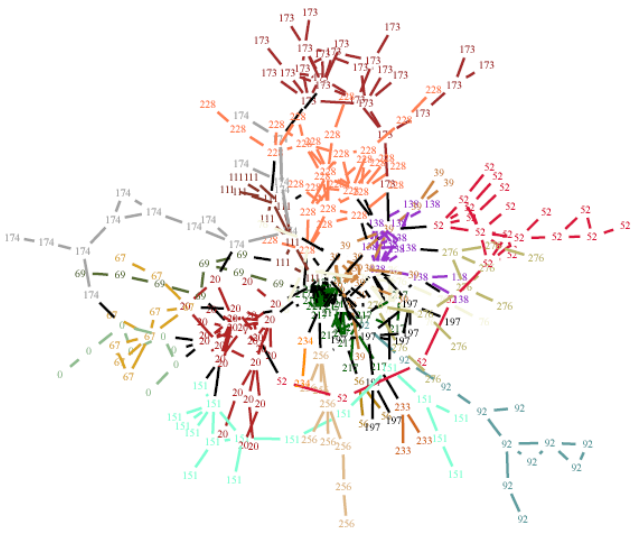

Figure 2: Network community assignments in the LCD bidding network, with colors and labels indicating nodal community membership

of each circle is proportional to the number of auctions in each community, and significant keywords are listed in order (with the most over-represented first).

For the Canon market with both edge filtering and edge weighting, there were 25 communities ranging in size from 3 to 339. Of these, 15 were major communities.[9] Twelve of the major communities had significant keywords, with a range of 1 to 6 significant keywords per community. The figure illustrates these 12 most significant communities. Ten of these communities had exactly one camera model keyword (sd500, a620, s2, a70, a520, s50, sd200, s80, s410, s110) and two communities had two camera keywords (sd400 and sd450; pro1 and g6). These are natural substitutes: the sd400 is the predecessor to the sd450, and the two camera models are extremely similar—they share the same resolution and memory card, with the major difference being a slightly larger screen on the sd450;[10] the pro1 and g6 are 2 of the 6 "high-end" Powershot digital cameras that Canon offers and both take the same type of compact flash memory card.[11]

LCD Screens

For the LCD market, we identified 42 communities ranging in size from 3 to 973. Of these, 12 were major communities. Ten of the major communities had significant keywords, with a range of 1 to 7 significant keywords per community; these 10 communities are shown in Figure 4. Eight of these 10 communities had a size (15", 17", 19", 20", 24") and/or

[6] In practice, it is possible to generate networks and community assignments where $Q < 0$. Furthermore, for any given network, it may be impossible to attain $Q = 1$. In earlier work, Newman (Newman 2003) normalized the value of Q; however, he argues that the unnormalized Q is more informative (Newman 2004).

[7] In fact, we found Q values of 0.72 in the LCD eBay market and 0.77 in the Canon eBay market—higher even than the modularity reported in the literature for Amazon.

[8] We use a simple statistical method to extract keywords to describe a collection of auctions (Jin 2006), looking for words that are both widely shared and over-represented, in the sense of TF/IDF methods (Salton & Buckley 1988).

[9] We define a *major community* as one with at least 1% of the total auctions in the maximally connected component.

[10] One review (www.dpreview.com/reviews/canonsd450/) termed the sd450 a "fairly minor upgrade" to the sd400.

[11] Information obtained from the official Canon Powershot website (www.powershot.com).

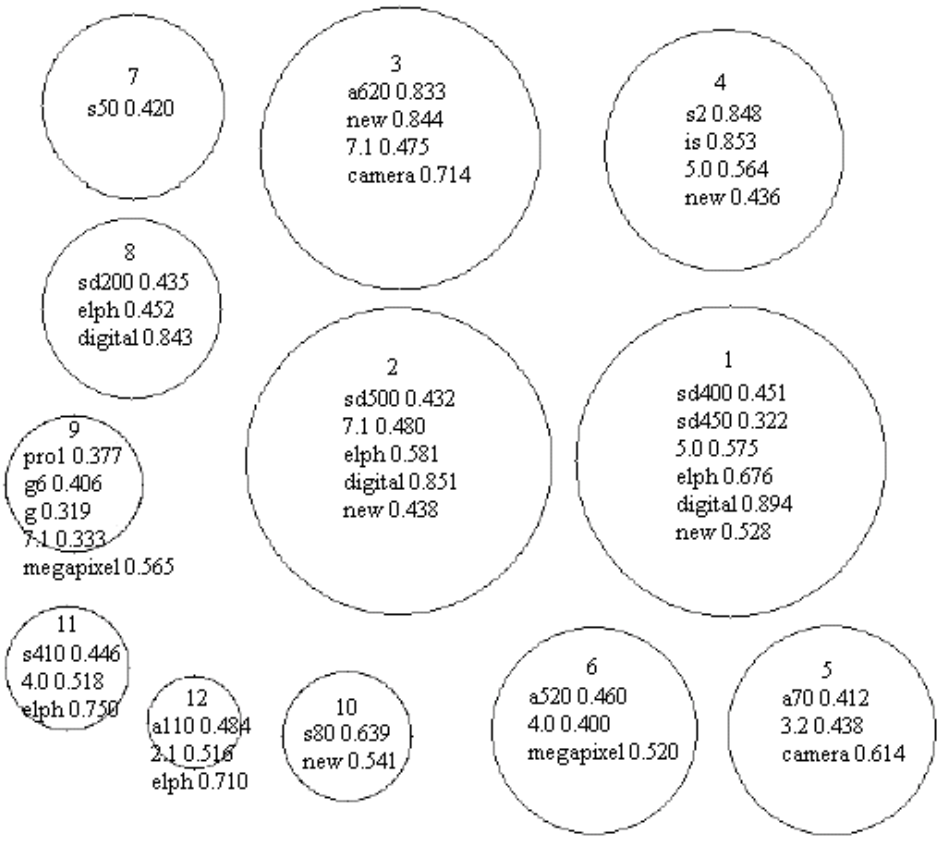

Figure 3: Proportionally sized Canon "substitutes" communities, with 0.8 edge thresholding and weighted edges

a specific model number (e173fp, 2005fpw, 2405fpw) as a keyword. Importantly, no communities had more than one size or more than one model number. This property is desirable because it is unlikely that different sizes or models of monitors serve as good substitutes.

The communities are not quite as distinct as in the Canon market—for example, there are two communities with e173fp as a keyword, and also two communities with 15" as a keyword. Thus, there is some evidence of oversegmentation. Nonetheless, when taken together, the keywords for the 12 major communities do seem to encapsulate the significant areas of the market—for example, all major monitor sizes are represented in at least one of the major communities.

An interesting anecdote is that examining the individual auctions in the LCD community corresponding to 24" 2405fpw Dell monitors (community 6 in Figure 4) revealed one case where our method correctly grouped an auction with similar other auctions even though the seller had listed it in an incorrect category. In this case, the seller listed the item in the 19-inch Dell monitor category, but our method nonetheless grouped it with other 24" monitors of the same model.

Results: Complements

With respect to complements goods, we assume that buying goods together is evidence of complementarity. Complementary goods can occur in sets larger than two, but as a starting point, we only consider pairs of goods.

To study complementarity, we collected two additional data sets. The first set consists of all auctions matching "Secure Digital" in the Secure Digital (SD) memory card category over a given time period. The second set consists of all auctions matching "Compact Flash" in the Compact Flash (CF) memory card category. (These additional data sets shared the same time period as our Canon data set.)

It seems reasonable that memory cards and cameras have complementary relationships. Furthermore, different models of camera require specific formats of memory cards—either CF or SD. Thus, we can assess the effectiveness of our methods by comparing the strength of complementary relationships detected for each camera model and the two types of memory cards. Out of the 12 significant camera communities from the previous section, 7 are SD cameras and 5 are CF cameras.

Our approach was two stage: first we identified (substitutes) communities of interest, and then constructed a new graph in which these communities of interest are nodes, and edges are weighted to indicate the number of distinct win-

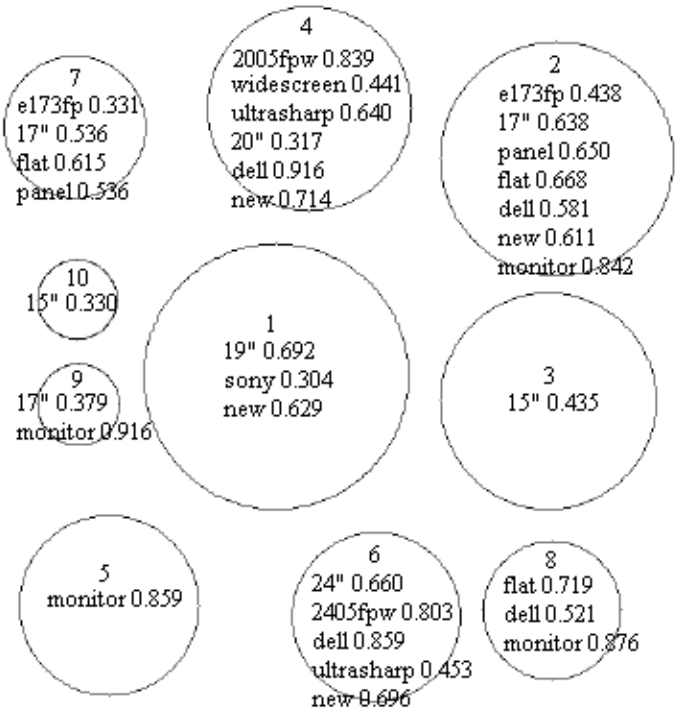

Figure 4: Proportionally sized LCD "substitutes" communities, with 0.8 edge thresholding and weighted edges

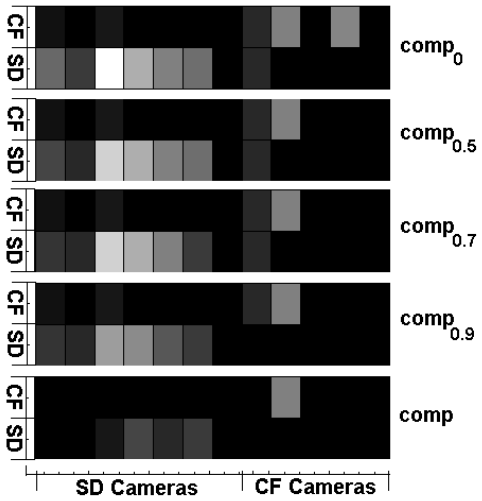

Figure 5: Density plot of complementarity between digital camera and memory card communities (SD: secure digital, CF: compact flash) in the Canon eBay market. Brighter boxes indicate higher complementarities

ners that submit competitive bids (according to some threshold fraction of the winning bid) in auctions that fall into the two communities.

We need a method to evaluate the strength of the complementary relationship between two communities of goods c_1 and c_2. Intuitively, if the two communities have a large number of shared winning bidders, then it is likely that they have high complementarity. We define $comp(c1, c2)$ as:

$$comp(c1, c2) = \max\{cpct(c1, c2), cpct(c2, c1)\} \quad (1)$$

where $cpct(a, b)$ is the number of distinct winning bidders in a that also win at least one auction in b, divided by the total number of distinct winning bidders in a. For cases where there are few shared winning bidders, we define $comp_T(c1, c2)$ as:

$$comp_T(c1, c2) = \max\{cpct_T(c1, c2), cpct_T(c2, c1)\} \quad (2)$$

where $cpct_T(a, b)$ is the number of distinct winning bidders in a that also place a bid that is at least a fraction T of the closing price of an auction in b, divided by the total number of distinct winning bidders in a. Note that $comp_1(c1, c2)$ is essentially equivalent to $comp(c1, c2)$—the only difference is that $comp_1(c1, c2)$ includes bids that tied the closing price but lost.

We examined $comp_T(c1, c2)$ for 14 communities, twelve of which were camera communities. In addition, we generated one CF community and one SD community. Each of the memory card communities contained a random 10% sample of successfully sold auctions in their respective markets.[12] Figure 5 displays the complement function values for the CF and SD memory card communities in relation to the camera communities. Most of the values for $comp_1(c1, c2)$ are 0—that is, there were no shared winners between most of the community pairs. For 4 of the 7 secure digital camera communities, there was a nonzero $comp_1(c1, c2)$ value between the camera community and the secure digital community. For 1 of the 5 compact flash camera communities, there was a nonzero $comp_1(c1, c2)$ value between the camera community and the compact flash community. There were no false positives.

We next assessed whether we could identify more associations by relaxing the threshold. At the least restrictive value of T, $comp_0(c1, c2)$, the function is able to associate 3 more camera communities with their correct memory card type. In the process, however, one compact flash camera community is potentially misclassified, since it has similar $comp_0(c1, CF)$ and $comp_0(c1, SD)$ values.

The data also suggest that there is some between-camera-community complementarity—that is, there are bidders who win multiple cameras from different communities. Figure 6 depicts the full 14×14 density plot for $comp_1(c1, c2)$. As we see, only 2 of the 4 camera communities with memory card $comp_1(c1, c2)$ relationships had their strongest $comp_1(c1, c2)$ value with the memory card community. The other two communities had stronger complement relationships with another camera community.

[12]We did not generate substitute communities for the memory card markets, as we simply wanted to test for complements between cameras and card types in general.

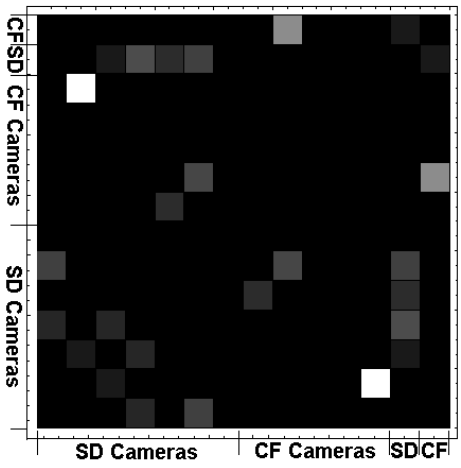

Figure 6: 14×14 Density plot for $comp_1(c1, c2)$ (brighter boxes indicate a higher value)

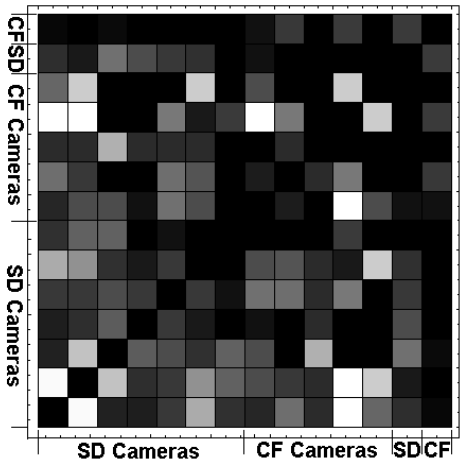

Figure 7: 14×14 Density plot for $comp_0(c1, c2)$ (brighter boxes indicate a higher value)

As we decrease T, the between-camera-community values of $comp_T(c1, c2)$ begin to dwarf the camera-memory card values of $comp_T(c1, c2)$. This fact becomes evident in Figure 7, which depicts a 14×14 density plot for $comp_0(c1, c2)$. While 8 of the 12 camera communities are associated with the correct memory card type, none of these relationships had the highest value of $comp_0(c1, c2)$ for their row. Thus, while relaxing $comp_T(c1, c2)$ can result in the identification of more associations, there are trade-offs in increased misclassification and noise.

Interestingly, stronger complementarity seems to exist between secure digital cameras and secure digital cards than between compact flash cameras and compact flash cards. Further investigation is needed to determine if this is due to an intrinsic property of the goods, or is rather a by-product of our choice of metric.

Looking Forward: Generative Modeling and Bayesian Inference

A criticism of the methods that we have explored in this work is that they are *ad hoc*. One first proposes a network to capture some underlying semantic property via community structure, and then applies a community detection algorithm without resort to an explicit model.

Following the recent directions of Kemp and Newman (Kemp *et al.* 2006; Newman & Leicht 2007), a more promising approach seems to be one of combining generative modeling with Bayesian inference. One defines a generative model that allows for particular relational structure between objects (e.g., between bidders and auctions) and is also stochastic. This in turn induces a graph whose edges, weights, and other parameters are considered to be random variables from an appropriately chosen probability distributions. A number of fitting methods are then available to estimate these parameters from observed data, as well as to compare and select from a variety of possible models.

For instance, as described in the citations above, one can posit a Dirichlet process prior model for community structure wherein the generative model captures: (a) an endogenous number of communities; (b) a distribution on the propensity of linking to items in other communities (and in the same community), conditioned on a particular community assignment; and (c) links between items given a model for linking propensity. Not only does such a scheme enable principled inference procedures that can also be extended to model the dynamic evolution of marketplaces and temporal dependencies within them, it also provides a means of uncertainty quantification for the resultant parameter estimates, an important consideration when scientific conclusions are being drawn from the data under study.

Summary

In this paper we have collected behavioral data from bidders in two eBay markets and from this data inferred aggregate properties about bidder preferences. The basic idea is that by their bidding behavior bidders indicate "revealed preference" information; e.g., two goods are natural substitutes (when auctions are part of a community in which bidders tend to win in only one auction) or two goods are natural complements (when auctions are part of a community then tend to share winning bidders). Although current community detection methods seem to be a reasonable tool we consider much of this methodology *ad hoc*: which graph to construct, how many communities to look for, how to interpret the results? As a future direction we have advocated the combination of generative modeling with Bayesian inference. We are currently pursuing this direction on new data sets.

References

Bapna, R. 2004. User heterogeneity and its impact on electronic auction market design: An empirical explanation. *MIS Quart.* 28:21–43.

Clauset, A.; Newman, M. E. J.; and Moore, C. 2004. Finding community structure in very large networks. *Phys. Rev. E* 70:066111.

Flake, G. W.; Lawrence, S.; Giles, C. L.; and Coetzee, F. M. 2002. Self-organization and identification of Web communities. *IEEE Comput.* 35:66–70.

Girvan, M., and Newman, M. E. J. 2002. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA* 99:7821–7826.

Hahn, J. 2001. The dynamics of mass online marketplaces: A case study of an online auction. In *Proc. SIGCHI Conf. Hum. Factor. Comput. Syst.*, 317–324.

Handcock, M. S.; Raftery, A. E.; and Tantrum, J. M. 2007. Model-based clustering for social networks. *J. R. Statist. Soc. A* 170:301–354.

Jin, R. K.-X. 2006. Leveraging bidder behavior to identify categories of substitutable and complementary goods on eBay. Senior thesis in computer science, Harvard College.

Kemp, C.; Tenenbaum, J. B.; Griffiths, T. L.; Yamada, T.; and Ueda, N. 2006. Learning systems of concepts with an infinite relational model. In *Proc. 21st Natl. Conf. Artif. Intell. (AAAI-06)*.

Kemp, C.; Griffiths, T. L.; and Tenenbaum, J. B. 2004. Discovering latent classes in relational data. Technical Report MIT-CSAIL-TR-2004-050, Massachusetts Institute of Technology, Cambridge, MA.

Lehmann, B.; Lehmann, D.; and Nisan, N. 2001. Combinatorial auctions with decreasing marginal utilities. In *Proc. 3rd ACM Conf. Electron. Commer.*, 18–28.

Lohse, G. L., and Spiller, P. 1998. Electronic shopping: The effect of customer interfaces on traffic and sales. *Commun. ACM* 41:81–87.

McGuinness, D. L. 2001. Ontologies and online commerce. *IEEE Intell. Sys.* 16:9–10.

Newman, M. E. J., and Leicht, E. A. 2007. Mixture models and exploratory data analysis in networks. *Proc. Natl. Acad. Sci. USA*. To appear.

Newman, M. E. J. 2003. Mixing patterns in networks. *Phys. Rev. E* 67:026126.

Newman, M. E. J. 2004. Fast algorithm for detecting community structure in networks. *Phys. Rev. E* 69:066133.

Palla, G.; Derényi, I.; Farkas, I.; and Vicsek, T. 2005. Uncovering the overlapping community structure of networks in nature and society. *Nature* 435:814–818.

Reichardt, J., and Bornholdt, S. 2004. Detecting fuzzy community structures in complex networks with a Potts model. *Phys. Rev. Lett.* 93:218701.

Reichardt, J., and Bornholdt, S. 2005. eBay users form stable groups of common interest. ArXiv Physics e-print physics/0503138.

Salton, G., and Buckley, C. 1988. Term-weighting approaches in automatic text retrieval. *Inform. Process. Manag.* 24:513–523.

Shah, H. S.; Joshi, N. R.; Sureka, A.; and Wurman, P. R. 2003. Mining eBay: Bidding strategies and shill detection. In Zaïane, O. R.; Srivastava, J.; Spiliopoulou, M.; and Masand, B., eds., *WEBKDD 2002 — Mining Web Data for Discovering Usage Patterns and Profiles*. Berlin: Springer-Verlag. 17–34.

Taskar, B.; Segal, E.; and Koller, D. 2001. Probabilistic classification and clustering in relational data. In *Proc. 17th Intl. Joint Conf. Artif. Intell. (IJCAI-01)*, 870–876.

Wasserman, S., and Faust, K. 1994. *Social Network Analysis: Methods and applications*. Cambridge, UK: Cambridge University Press.

Weiss, L. M.; Capozzi, M. M.; and Prusak, L. 2004. Learning from the Internet giants. *MIT Sloan Manage. Rev.* 45:79–84.

Yang, I.; Jeong, H.; Kahng, B.; and Barabási, A. L. 2003. Emerging behavior in electronic bidding. *Phys. Rev. E* 68:016102.