The Role of Chatting in Online Shopping

Tracy Chou
tracy.chou@cs.stanford.edu

Stephen Guo
sdguo@stanford.edu

Mengqiu Wang
mengqiu@stanford.edu

ABSTRACT

Our project sought to understand the behaviors within an integrated instant-messaging and e-commerce network, with the larger goal of understanding social influences on consumer behavior. We tackled this problem in three parts:

- *Statistics:* In the first part of our analysis, we aimed to gain a broad understanding of these networks by retrieving network statistics, such as the wealth distribution; node statistics, such as the relationship between chatting and shopping behavior for individual users; and community statistics, such as trade density within differently sized communities. For individuals, the amount of chatting was negatively correlated with purchasing activity but positively correlated with sales activity. Within communities, denser contact networks and greater messaging activity were correlated with higher trade volume.
- *Building Trust:* An important element on online sales is trust, since the possibility of deception is so much higher than for brick-and-mortar stores. We studied the phenomenon of "stickiness" (the tendency towards repeat business for individual users) as well as propagating trust through buyer-buyer interactions. For the latter, we isolated triads consisting of two buyers, each of whom had made purchases from the same seller, and investigated their messaging behavior. We found that buyers are very likely to purchase again from the same sellers, but sparsity of data on event triads did not give a strong signal on buyers propagating trust.
- *Signals Predicting Trade Likelihood:* To understand what different social influences affect trade likelihood, we performed a feature ablation study involving 30 different features, evaluating the metrics of average precision and area under the ROC curve. Among others, two of the most important signals were the total count of the buyer's outgoing messages, and the total count of the seller's previous transactions.

1. INTRODUCTION

It is widely accepted that there are strong social influences on consumer behavior, but without the appropriate datasets, it was previously not possible to do a large-scale empirical study of such network behavior. On the one hand, there has been research on social networks such as the MSN Messenger network [4]; on the other, there has been analysis of marketing and purchasing behaviors in e-commerce networks [3]. For our project, we have data from the world's largest consumer-to-consumer online marketplace in which users can list contacts as well as interact via both IM and sales activity. This provides us the opportunity to study the dynamics of three overlaid networks: a contact network, an instant-messaging network, and an e-commerce network.

One feature of our data is that it is all naturally observed, in contrast with data from controlled experiments such as that in "The Dynamics of Viral Marketing" by Leskovec et al. [3], where an incentive structure is constructed to study the effect of product recommendations in a social graph. An advantage of our approach is that we can see what network phenomena arise naturally, without the artifacts of artificially introduced interactions. However, we may have difficulty controlling for exogenous and other extraneous effects.

Our overarching goal for this project was to understand the social influences, as measured by features of the contact and IM networks, on trade activity in the e-commerce network.

2. THE DATASET

Our dataset comes from the world's largest consumer-to-consumer online marketplace, with approximately 150 million active users and transaction volume reaching nearly US$12 billion in the first half of 2009. The purchase network data is extremely rich in itself, but an even more unique aspect of this auction site is its integrated instant messaging network. Users on the site can message their friends, whom they have added to their contact lists, or also non-friends, such as sellers from whom they are considering purchasing some item.

To keep computation tractable, we restricted our analysis to a subsample of one million users. Starting from September 1, 2009, we recorded all trade activity for approximately two months (58 days), and then from this we extracted the first one million unique users who were either buyers or sellers. These constitute the nodes in our graphs.

For the contact network, we added an undirected edge between two users if they had listed each other as contacts during the time period. For the IM network, we added an undirected edge between two users if a message was exchanged between them during the time period. For the e-commerce

network, we added a directed edge for each buyer and seller pair for whom a transaction had occurred during the time period.

We verified that this sampling method produced realistic subgraphs throughout the course of our analysis – for example, the degree distributions of the three networks are consistent, and the best community sizes match empirical studies on similar networks [5].

3. NETWORK STATISTICS

Basic statistics on the number of (non-isolated) nodes and edges in these networks and their greatest connected components are listed in Table 1.

As might be expected, the contact network contains fewer nodes than the others, but it is much more densely connected, with an average degree per node of 9.67. The IM network contains more nodes, since users can message people with whom they are not friends. However, since it reflects actual activity between users, which is sparser than friend connections, the average degree in the IM network thus drops to 5.21. By construction, the e-commerce network has the most nodes, but since trade activity has a higher cost, it occurs far less frequently than IM activity. The average degree is only 1.34.

The degree distributions all follow power laws; they look quite similar in slope but differ slightly on intercept. See Figure 1 for log-log plots of degree distribution for each of the three networks.

The spending and wealth distributions also follow power laws, a well-known phenomenon in economics. See Figure 2.

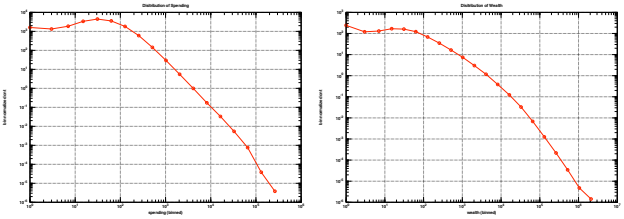


Figure 2: Spending distribution (left) and wealth distribution (right).

Analyzing trade volume broken down by transaction amount, we see that most transactions are within the 100 to 1000 RMB price range (approximately US$14 to $140). See Figure 3.

4. NODE STATISTICS

For a first pass at understanding the relationship between chatting and shopping on this website, we looked at node statistics relating to trade activity. Specifically, we drew scatterplots of number of contacts and messages versus sales and purchase volumes. Even binning produces somewhat noisy results, but it seems that number of contacts and messages both correlate negatively with purchase volume (see Figure 4) but positively with sales volume (see Figure 5). That is, successful sellers tend to be more active in the social graph, which may correspond to seeking potential customers. Buyers spend less the more they chat. One possible explanation is that chatting and shopping are competing uses of a buyer's time online.

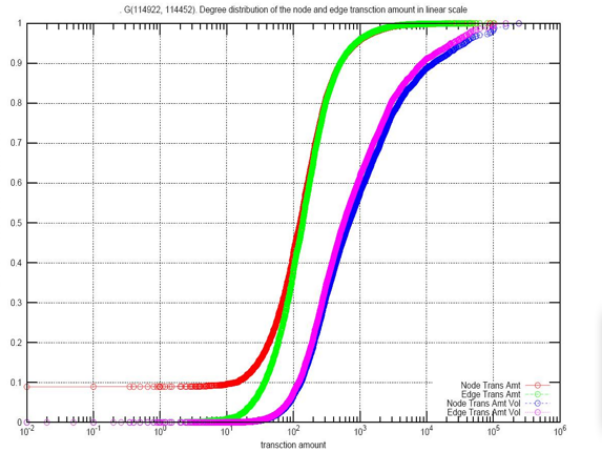


Figure 3: Trade volume broken down by transaction amount.

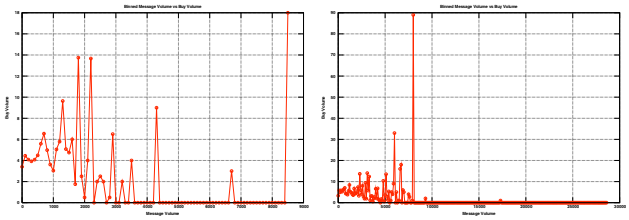


Figure 4: Purchase volume against number of contacts (left) and message volume (right), binned.

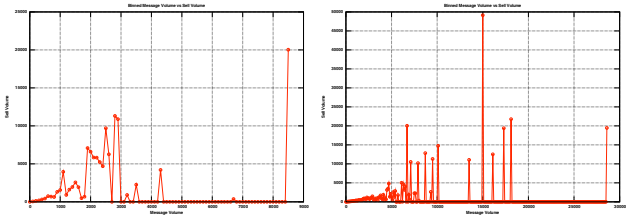


Figure 5: Sales volume against number of contacts (left) and message volume (right), binned.

5. COMMUNITY STATISTICS

Our next step was to study the communities within our three networks. We implemented the community finding algorithm proposed by Clauset et al. [2] in order to discover community structure within our networks. Appropriately for our dataset, the algorithm works best on sparse and hierarchical networks, where it runs in essentially linear time, $O(n \log^2 n)$ on n vertices; given the size of our network, no other community finding algorithms were tractable. We ran the algorithm on all three networks, but for brevity we will discuss the e-commerce network only.

The community partitions came out as expected. Most communities are very small, but there are a handful of very large

Network	Nodes	Edges	Avg Deg	Nodes in GCC	Edges in GCC	% Nodes in GCC	% Edges in GCC
Contact	663,346	6,416,086	9.67	661,491	6,414,068	99.72%	99.97%
IM	750,158	3,908,339	5.21	748,950	3,907,301	99.84%	99.97%
E-commerce	1,000,000	1,337,497	1.34	958,952	1,306,171	95.90%	97.66%

Table 1: Basic graph statistics.

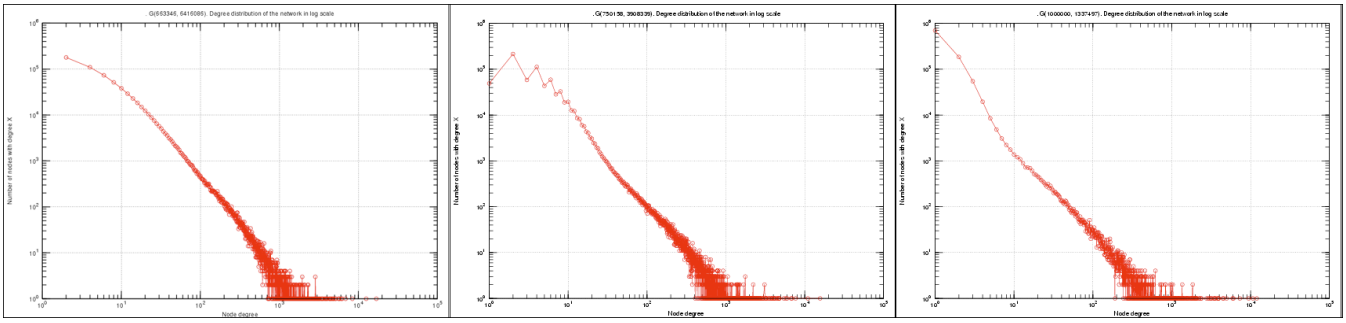


Figure 1: Degree distributions for the contact, IM, and e-commerce networks, respectively.

ones. Overall, there is an inverse relationship between community size and count of communities of that size, except for a notable bump at around size 100, which is akin to an "optimal" size of human communities [5]. See Figure 7.

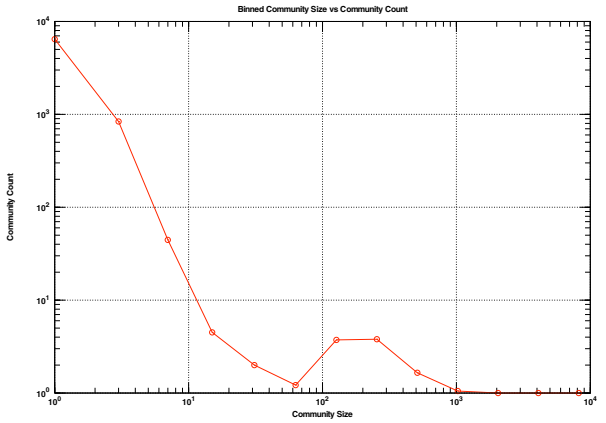


Figure 6: Count of communities of size k, versus k, in the e-commerce network.

Next, we compared community size with trade density, i.e. trade activity normalized over the size of the community. We found no significant variations due to community size; in fact, the amount of trade activity per node remained essentially constant for nodes in communities of all sizes. See Figure 7.

Disregarding community sizes, however, the community units do contain relevant information about activity of their members. We found a positive correlation between contact and message density with trade density in a community, suggesting that overall levels of social activity are an indicator of the amount of e-commerce. See Figure 8.

Lastly, with respect to communities, we sought to understand the relationships between the different networks. We picked out all communities from the e-commerce network of

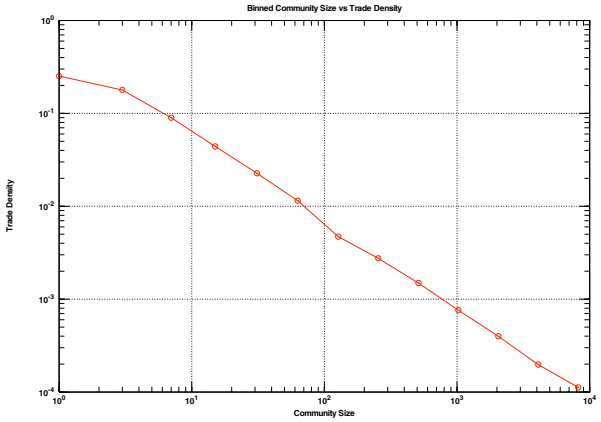


Figure 7: Trade density vs. community size in the e-commerce network.

significant size, which we defined as containing greater than 1000 nodes, and laid these out as "supernodes" in a Pajek visualization [1]. We then added edges between community nodes if the strength of the connection in terms of number of contacts, messages, or sales, was greater than 10 percent of the maximum strength. The e-commerce network has the most global edges defined in this way, while the contact network is more clustered; the message network is the sparsest but contains the same backbone as do the other graphs. See Figure 9.

6. REPEAT BUSINESS

Having retrieved general statistics on our three networks, we then moved on to the task of predicting trade activity. A well-studied effect is that of "stickiness" and customer loyalty [6], so we first examined the probability of first-time versus repeat business.

In our dataset, the baseline probability of a transaction occurring between a random buyer and seller, chosen from users who had bought or sold during our time period of anal-

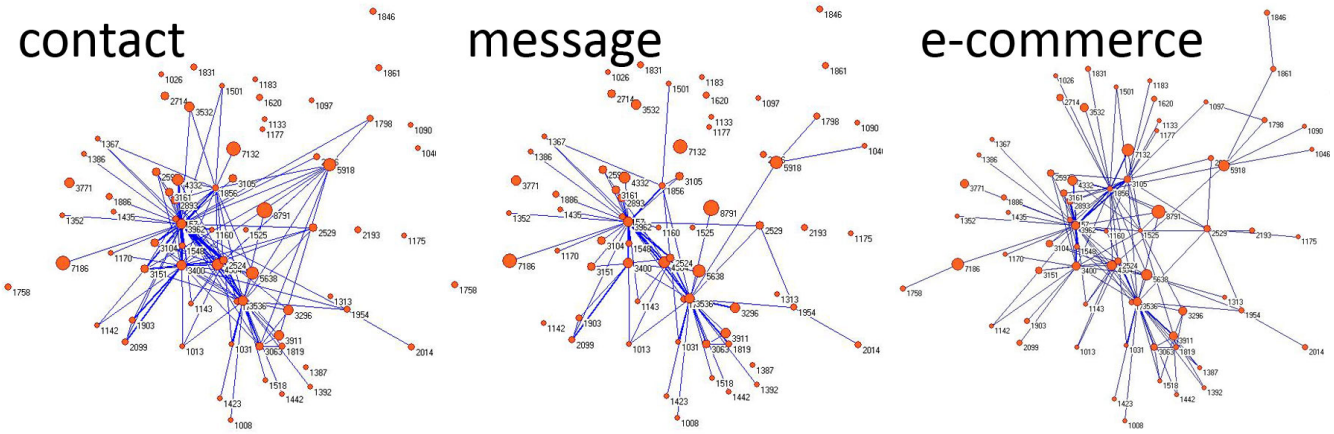


Figure 9: A comparison of cross-community activity in the e-commerce network.

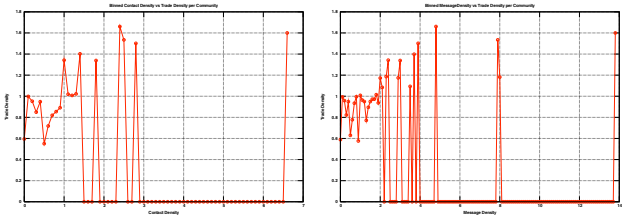


Figure 8: Community trade density against contact density (left) and message density (right), binned.

ysis, is extremely low: 0.000000025239, very close to zero. However, given that a transaction has already occurred between a buyer and seller, the probability of repeat business is quite high: 0.41431. In a large marketplace, especially one where trust is a difficult commodity to come by, it is not surprising that buyers would return to the same sellers for future purchases.

In accordance with business resesearch on similar subjects, we saw an exponential decay in repeat business over time, i.e. the stickiness effect wears off. See Figure 10.

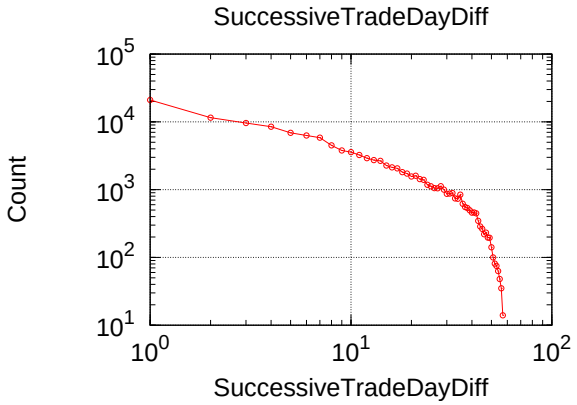


Figure 10: Count of repeat business k days after transaction, plotted against k.

7. EVENT SEQUENCES IN TRIADS

Each triad of interest in our analysis is composed of two buyers B_1 and B_2 who bought from the same seller S at times t_1 and t_2, respectively, with $t_1 < t_2$, and the seller S. For simplicity and ease of illustration, we only considered the first purchases made by B_1 and B_2, since the effect of repeat business was shown above.

We calculated empirical likelihoods of several events:

- B_1 and B_2 exchanged a message before t_1: 0.000018223162
- B_1 and B_2 exchanged a message between t_1 and t_2: 0.000017304410
- B_1 and B_2 exchanged a message after t_2: 0.000026855680
- B_1 and B_2 were contacts before t_1: 0.000039183829
- B_1 and B_2 became contacts between t_1 and t_2: 0.000005752512
- B_1 and B_2 became contacts after t_2: 0.000008521893

From these results, we see that the probability of B_1 and B_2 exchanging messages is about the same before and after the transaction at t_1 and before t_2. However, it is more likely for them to be in a triad together if B_1 and B_2 were contacts before t_1. That is, it is much more likely for a buyer to purchase from a seller if one of the buyer's contacts did so previously.

Although we had some interesting signals in this triad analysis, the sparsity and low quality of data made it difficult to do further extensions. Along those lines, we had originally planned to study the "price of trust", but the data was insufficient.

8. PREDICTING TRADE ACTIVITY

For this part of our project, we examined the effect of various network features on a maximum-entropy classifier for predicting trade activity. Our work used methodology similar to that in "User Grouping Behavior in Online Forums" by Shi et al. [7].

To create our dataset for this classification task, we defined each "trade event" to be a tuple consisting of a buyer, a seller, and a particular date. An event was positive if a transaction actually did take place between the buyer and seller on that date, and negative otherwise. We first included all 3,024,629 positive events that were observed in our 58 days of data. For each of these events, we generated a negative event for the same buyer and seller pair, using a randomly sampled date on which they did not trade (unless they traded every day in our observation period). Then we sampled 3000 buyers and 3000 sellers, and for each of the 9,000,000 possible pairings, we randomly sampled a date on which they did not trade (unless they traded every day), to generate negative events. At the end of this process, we had 3,024,629 positive events total and 12,236,549 negative events total.

We started with 30 features included in the classifier, and then removed them one at a time to compare the effect on average precision (AP) and area under the ROC curve (AUC).

Some of the more interesting boolean features included whether or not the buyer and seller had traded before, which boosted the output prediction of a trade from 0.000000000000 to 0.659876942635; whether or not the buyer and seller were in the same community, which increased the likelihood of trade from 0.075542218983 to 0.499092608690; if the buyer and seller were already contacts, which raised the likelihood from 0.166977763176 to 0.571738481522; if the buyer and seller had mutual contacts, which boosted the likelihood from 0.192327067256 to 0.474116921425; and finally, if the buyer and seller had messaged each other before, which increased the likelihood from 0.117516040802 to 0.626082539558.

The two most significant real-valued features for the maximum-entropy classifier were the number of the buyer's outgoing messages and the total volume of the seller's previous trades. The fact that the buyer's outgoing message activity correlates so strongly with the likelihood of trade is particularly interesting given that our earlier results showed that overall chatting activity, including both incoming and outgoing messages, is negatively correlated with purchasing activity. One possible explanation is that a large number of outgoing messages indicates that a user is seeking out information or terms for purchase. The fact that the seller's previous trade activity is so important points to the "rich-get-richer" effect. One feature that surprisingly had no relevance was the number of mutual contacts between the buyer and seller. Some of the other features are graphed in Figure 11.

All results are summarized in Table 2.

9. SUMMARY

Over an integrated instant-messaging and e-commerce network, we studied the effects of the social graph and social activities on trade activity.

For computational tractability, we sampled one million users of the 150 million total to be the nodes of our graphs. We then constructed three graphs from this nodeset, with the edges derived from the contact network, the instant message network, and the trade network.

We verified previously known phenomena like the power law distribution of wealth, the "rich-get-richer" effect for sellers, and the "stickiness" effect of repeat business.

We found that the number of contacts and level of messaging activity negatively correlated with purchasing activity but positively correlated with sales activity. This suggests that for buyers, chatting and shopping are competing activities, but for sellers, maintaining a strong social network correlates with better business. However, we also discovered that the number of outgoing messages that a buyer sends is positively correlated with the probability of a trade, perhaps reflective of information or deal seeking behavior.

Within communities, greater contact network density and level of chatting activity correlate positively with higher levels of trade activity. This suggests that users in communities that are bound more tightly by social interaction are also more likely to buy and sell from each other.

To study the effect of a wide range of social features on the probability of trade activity, we performed a feature ablation study using a maximum-entropy classifier. Many of the results confirm intuition – for example, that the number of purchases a user had made previously was strongly correlated with the probability of making another purchase. However, one surprising result was that the number of mutual contacts between a potential buyer and seller did not effect the probability of the transaction occurring.

Future directions would incorporate our findings in this project into a more comprehensive network model.

10. REFERENCES

[1] V. Batagelj and A. Mrvar. Pajek - program for large network analysis. *Connections*, 21:47–57, 1998.

[2] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. Aug 2004.

[3] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing, Sep 2005.

[4] J. Leskovec and E. Horvitz. Worldwide buzz: Planetary-scale views on an instant-messaging network - microsoft research.

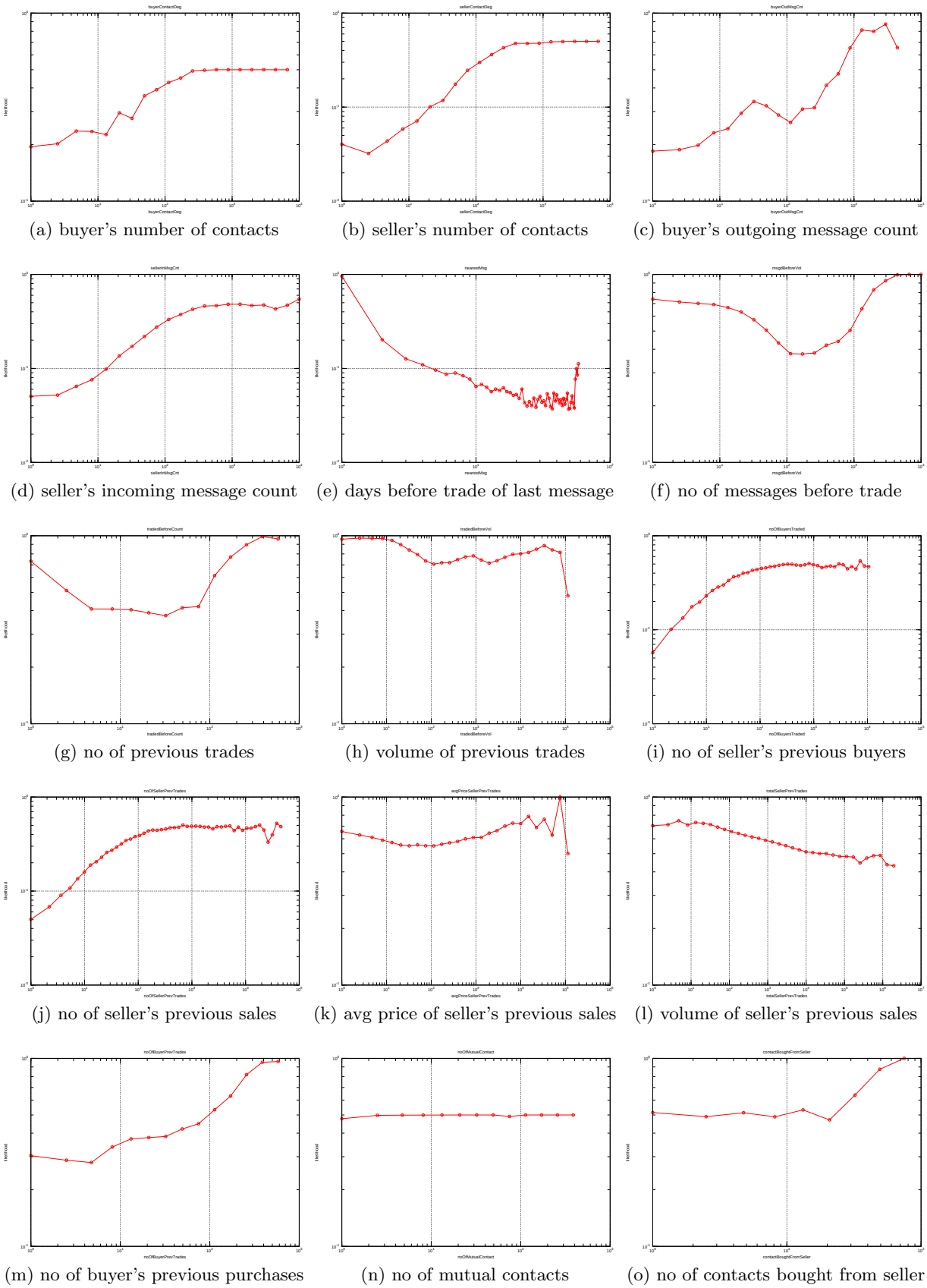
[5] J. Leskovec, K. J. Lang, A. Dasgupta, and M. W. Mahoney. Statistical properties of community structure in large social and information networks. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 695–704, New York, NY, USA, 2008. ACM.

[6] F. F. Reichheld and P. Schefter. E-loyalty. *Harvard Business Review*, 78:105–113, 2000.

[7] X. Shi, J. Zhu, R. Cai, and L. Zhang. User grouping behavior in online forums. In *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 777–786, New York, NY, USA, 2009. ACM.

	Average Precision	Area Under Curve
incl. all features	0.22182565535	0.834181558022
excl. 00: no of contacts the buyer has	0.22182565535	0.834181558022
excl. 01: no of contacts the seller has	0.376797391727	0.627996880303
excl. 02: bool, if the buyer and seller do not have mutual contacts	0.22182565535	0.834181558022
excl. 03: bool, if the buyer and seller are contacts prior to transaction	0.230614447634	0.834182400589
excl. 04: number of buyer's outgoing messages	**0.0158543685653**	**0.629804548707**
excl. 05: no of seller's incoming messages	0.143366731142	0.729499994981
excl. 06: no of days prior to transaction buyer, seller last exchanged a message	0.225685849648	0.83392429371
excl. 07: no of conversations between buyer, seller before transaction	0.22922928851	0.771186121317
excl. 08: no of messages prior between buyer, seller before transaction	0.225051833446	0.834183877653
excl. 09: bool, if buyer and seller had conversation previously	0.22183781155	0.834181556654
excl. 10: no of previous transactions between buyer, seller	0.227821698938	0.834181339277
excl. 11: total previous trade volume between buyer, seller	0.231460599036	0.833909365159
excl. 12: avg price of previous transactions between buyer, seller	0.240846515724	0.744943268054
excl. 13: bool, if the buyer and seller traded together previously	0.221837828068	0.834181552195
excl. 14: no of unique buyers the seller traded with previously	0.206627853943	0.834748555546
excl. 15: no of transactions seller has engaged in previously	0.215706610199	0.751039242372
excl. 16: avg price of seller's previous transactions	0.382465253867	0.666553949828
excl. 17: total volume of seller's previous trades	**0.0619264205711**	**0.562017600675**
excl. 18: no of unique sellers the buyer traded with previously	0.41772668624	0.676440960522
excl. 19: no of transactions buyer has engaged in previously	0.417627316574	0.676242212393
excl. 20: avg price of buyer's previous transactions	0.422065876791	0.683570884149
excl. 21: total volume of buyer's previous trades	0.446741949898	0.750541679922
excl. 22: bool, if the buyer and seller are in the same community	0.417729595903	0.676446557478
excl. 23: bool, if the buyer and seller are not contacts (inv 3)	0.417728544781	0.676445563154
excl. 24: bool, if buyer and seller never had conversation previously (inv 9)	0.417728240948	0.67644444479
excl. 25: bool, if the buyer and seller never traded together previously (inv 13)	0.417728446796	0.676444274936
excl. 26: bool, if the buyer and seller are not in the same community (inv 22)	0.417728535282	0.676443818998
excl. 27: the number of mutual contacts between the buyer and seller	0.41773040996	0.6764486321
excl. 28: bool, if the buyer and seller do not have mutual contacts (inv 2)	0.417729599233	0.676446565017
excl. 29: the number of the buyer's contacts who have bought from the seller	0.417730402612	0.676448616959

Table 2: Maximum-entropy classifier prediction results.



(a) buyer's number of contacts (b) seller's number of contacts (c) buyer's outgoing message count

(d) seller's incoming message count (e) days before trade of last message (f) no of messages before trade

(g) no of previous trades (h) volume of previous trades (i) no of seller's previous buyers

(j) no of seller's previous sales (k) avg price of seller's previous sales (l) volume of seller's previous sales

(m) no of buyer's previous purchases (n) no of mutual contacts (o) no of contacts bought from seller

Figure 11: Likelihood plotted against various non-binary features used in classification.