Pairwise choice as a simple and robust method for inferring ranking data

Alexander Peysakhovich
Facebook
alexpeys@fb.com

Virot Chiraphadhanakul
Facebook
virot@fb.com

Michael Bailey
Facebook
mcbailey@fb.com

ABSTRACT

One of the largest challenges for a recommender system is building a ranking of "quality" or "relevance" in situations where these features cannot be observed directly. These models are often trained on various types of survey data, including Likert-scale quality ratings or pairwise comparison surveys, but there has been little work detailing the efficiency of these techniques for eliciting quality ranking and a parsity of work on how to analyze and interpret pairwise choice data. We present techniques for using pairwise choice data for quality ranking and we find, under simulation, that Likert scale elicitation is more efficient under the best possible conditions but in the presence of differential item functionality (i.e., the fact that different scale points may mean different things to different people) or low quality inputs (e.g., lack of attention or understanding by survey participants or noisily measured input features) pairwise comparison becomes a more efficient survey method. We confirm this finding by using different survey techniques to infer the relevance of individuals' Facebook News Feed stories. Pairwise choice elicitation can be finished quickly by survey participants, is easily to implement and scale, produces models with interpretable results and is robust to noise and interpretational issues. Thus, we argue, pairwise choice surveys have wide potential for application.

Categories and Subject Descriptors

J.4. [**Social and Behavioral Sciences]** Economics

Keywords

Recommender systems, crowd sourcing, behavioral economics

1. INTRODUCTION

The Internet is full of content, mostly videos and pictures of cats [19]. However, not every piece of content (and certainly not every cat video) is created equal and thus algorithmic sorting by relevance is an important part of many websites' business models.

Often, explicit user reviews make the sorting process easier (e.g., in the case of Netflix or Amazon streaming video) and much research has gone into recommender systems based on collaborative filtering or other approaches [17, 18]. However, in many other important cases eliciting individual ratings of each piece of content is either costly or downright impossible (imagine asking all Facebook or Twitter users to evaluate every story/tweet they see or redditors to rate every meme in the known universe

Our proposed method is as follows: a content platform (e.g. Facebook News Feed) samples a representative subset of individuals and presents them with pairs of randomly chosen content. For each pair individuals choose the more relevant of the two.

We show how to apply *discrete choice random utility models* to take the resulting survey data to estimate a ranking function. These models have a rich history in behavioral economics, psychology and game theory [12, 13]. They have been used to model phenomena from psychophysics (perception of sound/brightness etc…) [20], to estimating individual preferences for risk [22], to models of human learning [23] to larger models of markets where random utility functions are aggregated to build consumer demand [10]. As we will see, they are a natural fit for the problem of building ranking.

Our main results concern the robustness properties of this model to real world confounds such as inattention/lack of understanding by survey participants, noise in the measured features and heterogeneity in preferences. We also compare the pairwise choice elicitation mechanism to another commonly used elicitation mechanism: having individuals rate the quality or relevance of items one at a time using Likert scales.

Submitted for publication at WWW 2015 Conference Proceedings.

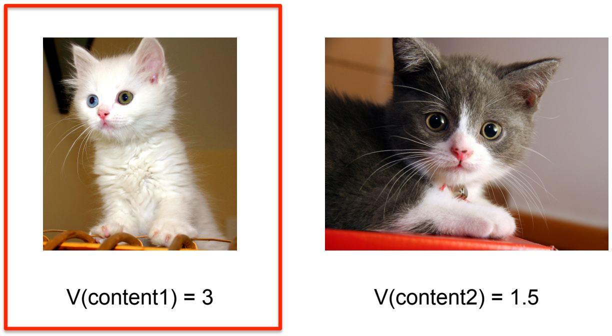

Figure 1: Individuals are shown 2 pieces of content and are asked to choose the higher quality one. Our assumptions are that each piece of content has an underlying intrinsic quality that individuals perceive with noise and choices reflect this noisy utility. We then use a logistic random utility model to recover true underlying mappings from features to quality.

Another solution to the relevance sorting problem comes in the form of voting schemes popularized by websites like reddit. These techniques employ the power of the crowd to find relevant content, however they have two disadvantages. First, they do not leverage the information that websites often have about a piece of content (attributes such as topics, features of the content author, similarity to other high-quality content, etc…). Second they are vulnerable to "herding" behavior where content can become highly ranked due to path-dependence (e.g., an early upvote) rather than actual quality [14, 15]. The pairwise choice method avoids these concerns.

2. METHODOLOGY

We now describe how we go from sets of pairwise choices to quality rankings. Our workhorse model is a simple utility model. Individuals are presented with two pieces of content. We index content by i and assume that each piece of content has an associated feature vector representation $\mathbf{f}_i$. For the purposes of the pairwise choice problem, let's call the items c_{left} and c_{right}.

Each individual, who we index by j, has a function $U_j\,(c_i)$ that map features of each item to its relevance or quality. We call these the users' *true utility functions.* When asked to indicate which of two items is higher quality individuals estimate this quality level but with some error.

This error around the true utility is used as a modeling assumption for many important psychological abstractions: human perception is itself noisy [20] and momentary emotional/physiological states can affect judgments [21]. In addition, this noise also serves as a modeling proxy for features that are not explicitly modeled/measured by the platform or analyst.

We refer to this noisy estimation as the individual's *perceived utility* $V_j(c_i)$. For the purposes of the simulations we assume the noise is normally distributed with mean 0 and variance s_e.

$$V_j(c_i) = U_j(c_i) + e$$

Further, we assume that the functions are linear in feature vectors so we get:[1]

$$U_j\,(c_i) = \boldsymbol{f_i}{*}\boldsymbol{B_j}$$

Finally, we assume that that the individual level utility weights $\boldsymbol{B_j}$ are drawn from a multivariate normal population distribution with a grand mean $\boldsymbol{B}_M$ and covariance matrix $\boldsymbol{S}_B$.

The process of choice then goes as follows: individuals are presented with items c_{left} and c_{right} (with feature vectors known to the analyst) and they estimate $V_j(c_{left})$ and $V_j(c_{right})$. They then make the choice that reflects their highest *perceived utility.* Figure 1 summarizes the process.

From a sample of individual choices our task is to recover and evaluate a population ranking function (here, a vector of weights $\boldsymbol{R}$). In other words, we seek to estimate the grand mean $\boldsymbol{B}_M$.[2]

Note that due to the assumptions we have made the probability of choosing an item i as being more relevant than an item k is proportional to the true utility difference of the two items. For this reason to estimate $\boldsymbol{R}$ we simply run a logistic regression with y

[1] Note that this assumption is not so restrictive. We could always take a basis expansion of the original set of features to approximate any non-linear function. Alternatively, we can think of our task as finding the linear function that best approximates the users' utility functions. We also note that the distribution assumptions on $\boldsymbol{B}$ and e are not restrictive and in fact the logistic random utility model we use is, in fact, mis-specified given our simulated data.

[2] An interesting and important extension of this methodology is to use it to build not just a ranking function that works on average, but to build individual-level ranking models. The focus of this paper is not to discuss this, rather it is to show the robustness and power of the pairwise choice methodology in the simplest case.

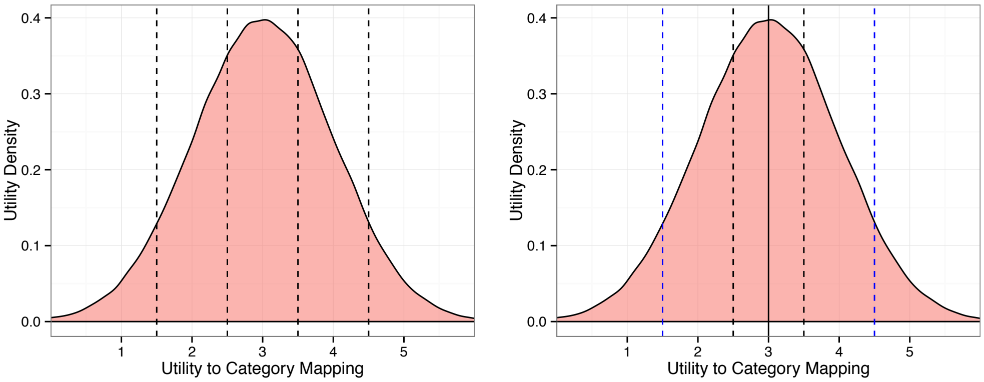

Figure 2: We model Likert scale responses as thresholds which define what quality values are, for a given individual, associated with a scale response. In the base model (left panel) all individuals use the same cutoffs which are in terms of standard deviations of the underlying uility function. To model DIF (right panel) we allow 'selectiveness' (which we model by the span of the blue lines), inner thresholds (black lines), and positivity/negativity bias (offset of all lines from central black line) to vary.

= "did the user choose the item on the left?" as our binary outcome and the difference in features as our independent variable. Formally the model is:

$$y_{ik} = logit((\boldsymbol{f_i} – \boldsymbol{f_k})\boldsymbol{R})$$

We use the recovered **R** as our ranking function.

The rest of this paper investigates two questions using simulations and real data:

1) How well does this procedure work?
2) How does pairwise choice compare to Likert-scale rating, another standard method of survey evaluation?

3. COMPARISONS TO LIKERT SCALES

One could also ask the user to rate the quality of the item on a continuous scale. In many cases the continuous version is quite difficult as quality or relevance may have no natural scale. Therefore, most surveys ask individuals to rate each piece of content on a Likert scale (e.g. from "1-Very Poor" to "5-Excellent").

On the surface these methods appear to be far more powerful than pairwise comparison. Out of n continuous valued answers we could build $(n^2 – n) / 2$ pairwise comparisons. In addition, because quality levels are now directly elicited rather than come out of a model, it is somewhat more straightforward to interpret the data. However, the next set of simulations will show that under ideal conditions the comparison advantage is modest at best and additional complications arising from the analysis of Likert scale data make pairwise comparisons more robust and simpler to use.

We employ simulation methods to investigate the relative power of each approach. Simulation methods have the advantage that they allow us to vary parameters explicitly as well as giving us a ground-truth to compare against.

Our simulations include P individual taking both pairwise comparison surveys and Likert scale surveys. Each person is presented with 10 questions. We generate each item as a feature vector with f dimensions (which we will vary) and we draw a utility function as described in the methodology section setting $\boldsymbol{B}_M$ to be a vector of 1's and the covariance matrix to have all feature weights be independent draws with identical variance s_B (which we will vary).

To generate Likert scale responses we assume that individuals have cutoffs (l_1, l_2, l_3, l_4) such that individuals give the item a 1 if its perceived relevance falls between *–Infinity* and l_1, a 2 if it falls in the interval $[l_1, l_2]$ and so on (see Figure 2 for visual demonstration).

Note that this assumes that every individual uses the same cutoff rating. This gives us a simple latent variable model that can be estimated using an ordered Probit regression [9].

We simulate P individuals each making 10 survey decisions and ask how well each method performs as a function of the model parameters. In particular we focus on two error metrics: first, we ask how well does the model recover the true mean utility vector in the population? That is, how close is the ranking function that we would recover (and presumably later use) to the true ranking function in the population?

Note that our question is subtler than it appears. Due to noise in our data we always expect to have

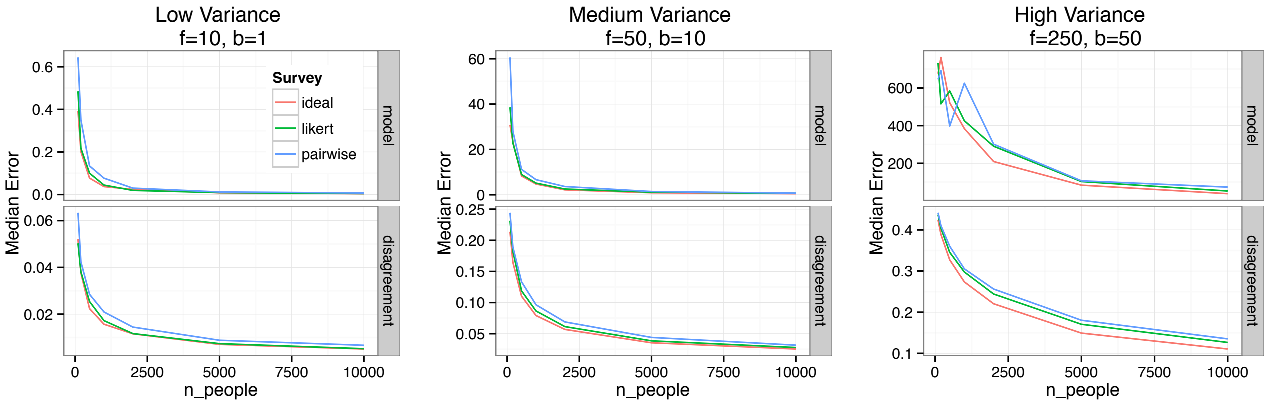

Figure 3: All methods converge to identifying the optimal grand mean ranking (with high variance it might take thousands of raters). Although Likert scale based rating outperform the pairwise comparison method, it does so only modestly and only under the ideal conditions of identical thresholds.

attenuation problems driving our regression coefficients towards 0 (that is, if we assume that both features and rankings are measured with error than any correlations that we infer between the measured values will be weaker than true population level correlations).

But, also note that we are not interested in recovering an *absolute value* but rather a ranking. This means we are interested not in how accurate the *absolute coefficients* are but rather in the *ratios* of weights (often called the *marginal rates of substitution*). In other words, we don't care about the absolute "weight" of a feature but rather how much one feature is worth relative to others in terms of adding quality.

As a quick example, suppose that the true mean weights are (1,1), then a model that recovers the weights (.5, .5) will have the wrong coefficients but will still rank all items correctly.

For our first metric then, we will consider how well each procedure allows us to recover the marginal rates of substitution, or, more formally we will use the distance metric given by

$$\sum_{k\,!=\,1} ((R_k/R_1)– 1)^2$$

Second, since we are interested in models that rank content we ask: if we were to randomly draw two pieces of content, how often would the recovered model correctly rank them relative to the ideal ranking given by using the true mean utility vector?[3]

For our simulations we vary:

1) the number of individuals recruited P: (n people)
2) the number of features whose weights need to be estimated: (f)
3) heterogeneity in preferences (i.e., the variance s_B): (b)

We fix the noise level (variance of e) to be equal to the sum of the variance of the features. In essence, we assume that a full 50% of the signal contained in the perceived value of each item is noise.

We run our simulations 100 times for each parameter value and take the median error metric (we take the median because the large variance cases sometimes cause large deviations in coefficient estimates, in practice we recommend employing some form of regularization in the logistic/probit regressions used to estimate MRS).

Figure 3 shows the results of our simulations. We see that as we increase P the ranking model estimated from pairwise choices quickly converges to be nearly identical to the model that would be estimated from our 5-point Likert scale. In addition, both models are very close to the performance of an ideal model where individuals are able to state their perceived utility for each item directly.

This is true (Fig 3, panel C) even when there is a very large feature number of features, very large heterogeneity in preferences (s_B = 50) and large noise. Even 10,000 survey participants is not enough to reach accuracy levels of above 90% in the

[3] We note that there are more complex ways of evaluating a ranking (e.g. metrics like NDCG [6] or Kendall's Tau [5]). Using these other evaluations would change the absolute numbers but not the flavor of our results. For the purposes of this exercise we use the pairwise transposition score because it is simpler for exposition.

transposition score. The stark difference between panels C and B suggests that a first stage of feature compression may be another important step to take when building ranking models.

Although the number of survey participants may look large, one should keep in mind that the surveys here are only 10 questions long and don't require significant setup and so can be administered by a content platform without requiring more than a minute or two of a user's time.

Of course these results beg the question: why use pairwise comparison surveys when a 5-point Likert scale is much more efficient at small samples and continues to be slightly more efficient with larger samples?

The main issue here, as with other applications of continuous scales in surveys, is that scale points (especially ones labeled with arbitrary values) can mean different things to different people. This is often termed *differential item functionality* (DIF) [3]. In our model, this corresponds to adding in the assumption that the "breakpoints" individuals use to assign a perceived utility level to a Likert scale category are heterogeneous across people.

We model the degree of this heterogeneity as follows: we assume that everyone has a baseline breakpoint as in the single threshold model and we jitter those thresholds. This is illustrated in Figure 2B. First, we set the "width" of the cutoff zones for 1 and 5. This can be thought of as different individuals having different cutoffs for when they rate something as "absolutely terrible" or "absolutely amazing." This is represented in Figure 2B by the blue lines. This gives us our first parameter *width* which is 3 standard deviations in the single threshold model.

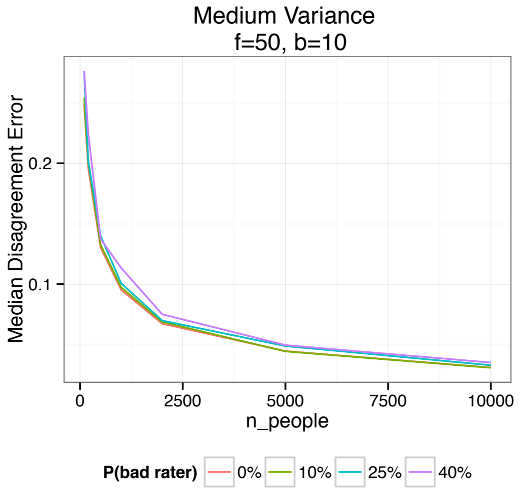

Figure 4: Pairwise choice elicitation is robust to the addition of even a large proportion of individuals who don't follow the survey directions and choose randomly.

In addition, we randomize the position of the inner thresholds (black lines). Finally, we also model individual level bias towards one or another side of the scale by offsetting the position of all thresholds by a constant (in the single threshold model the offset is 0, as demonstrated by the black line in Fig. 2B).

We now consider what happens when we add DIF to our simulations. We consider the medium variance case, but results are qualitatively identical for low and high variance parameter sets.

Re-running our simulations (Figure 4) allowing for heterogeneous thresholds shows that the presence of even mild DIF gives the pairwise choice model an advantage.

We do note that techniques exist to attempt to control for DIF including using a common anchor [4], estimating heterogeneous thresholds [1] and ensuring via training or strict instructions (i.e. "rubric grading") that individuals use identical thresholds. When these techniques are available, Likert scales may indeed be the most useful technique for survey design. However, for easy and simple surveys which gather power from pooling subject responses to estimate a mean parameter we believe that pairwise choice is a natural and robust surveying method.

We now consider another source of trouble for survey methods: lack of attention or interest by participants. What happens to our results if a substantial portion of

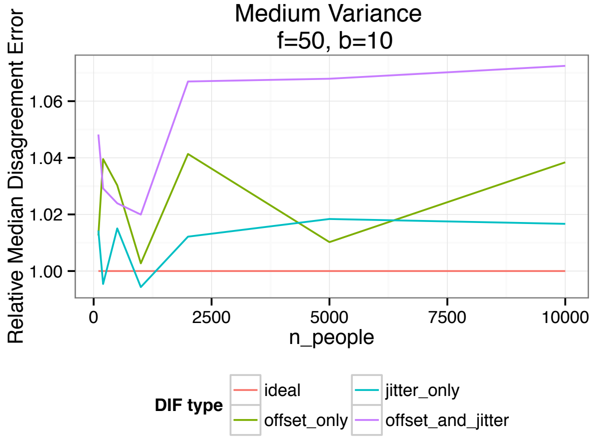

Figure 5: Adding DIF decreases the accuracy of Likert-elicited ranking functions.

participants are simply not performing the task as requested?

We add this to our simulations by considering that a portion of individuals can be "bad survey takers." We assume that a proportion *b* of individuals, instead of choosing to maximize their perceived utility simply choose randomly.[4] Figure 5 shows the results: the addition of even a large proportion of "bad survey takers" does not change our results very much.

4. EMPIRICAL EVALUATION

So far we have focused on simulation results. We now turn to the performance of these measures out in the wild. To do so, we look at data from Facebook surveys. In the summer and fall of 2014, we asked a random sample of Facebook users to rate the quality of Facebook News Feed posts.

There were two separate surveys, in one survey individuals were presented with randomly sampled stories generated by their friends or pages they were connected to and asked: "How much would you like to see this story in your News Feed?" which they answered using a 1-5 Likert scale.

In the other survey individuals were presented with two stories (also drawn randomly from the set of stories created by their friends or pages they were linked to) and asked to "[c]hoose the story that is most interesting to you." Both surveys had a small introduction section and took comparable lengths of time to complete.

We now want to compare which survey method elicits ranking data more efficiently. However, unlike in our simulations, we lack any notion of ground truth (if we knew which stories were most relevant, we wouldn't need to be doing these surveys).

Thus, we turn to a proxy metric to evaluate the effectiveness of our survey techniques: the variance of the estimated MRSs.

The intuition behind this choice of metric is as follows: both methods should, in theory, be consistent and thus estimate the same ranking model in the limit of infinite data. Thus, the question of "how well will this method estimate a ranking model with a sample size of *N*?" can be thought of as asking "how close are we to the *N=infinity* limit?" In other words: how much variation do we see in the estimated coefficients?

We estimate this as follows: we take a large sample of survey participants and calculate the standard error of the MRS for the sample by bootstrapping (using 200 bootstrap replicates) at the individual level (because decisions are correlated at the individual level). We then take the average standard error for each *N* participants over 240 runs of the simulation.

If this procedure sounds confusing, a simple intuition for it is that we are simply trying to replicate the procedure for the simulations used earlier in the paper but this time using real data and a different error metric.

We use a very simplified feature space for this exercise. For each story/individual pair we use the predicted probability that an individual will Like, Comment, Share or Click (when applicable) on this content. These probabilities themselves are trained from higher dimensional feature spaces including past interaction history of individuals with content similar to the content they are viewing. For a more detailed description of how these models operate see [24].

We note that our focus here is not on the input features, but rather on the precision with which both survey procedures estimate the ranking function which uses these features to estimate a ranking. We thus turn to these results plotted in Figure 6.

We see that in the real data sets the pairwise comparison-based survey is more efficient (can estimate the parameters of the ranking function more precisely) as a function of the inputs.

5. CONCLUSION

We have studied the problem of using 'small data' in

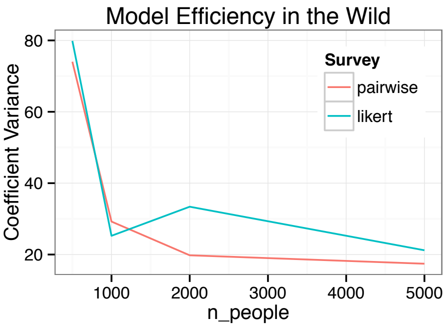

Figure 6: In a real experiment the pairwise choice-based procedure is able to estimate ranking functions more efficiently than the Likert-scale based procedure.

[4] Note that we could make other assumptions about the form of the behavior of bad raters (for example: they could have a bias towards the left or the right, rather than choosing randomly). Most of these would not change the results as instead of being picked up by the error term in the logit model, they would be picked up by the intercept.

the form of surveys in combination with ‘big data’ in the form of detailed feature knowledge to build models to rank content in terms of quality or relevance.

Our initial simulation and empirical results are promising: pairwise choice elicitation is a robust and scalable method for learning individual preferences. Likert-scale elicitation is likewise useful but when conditions are not ideal (e.g. there is DIF in the population) pairwise choice outperforms the Likert scale.

These survey methods have much theoretical interest, but at the end of the day they are inputs to an engineering enterprise and so more work needs to be done on evaluating the robustness of these methods in the field as well as the quality of the resulting ranking mechanisms.

6. ACKNOWLEDGEMENTS

We thank members of the Core Data Science team at Facebook, the News Feed team and the UX Research team for thoughtful feedback and discussion at all stages of this project. All errors remain our own.

7. REFERENCES

[1] F. Peracchi and C. Rossetti, “The heterogeneous thresholds ordered response model: Identification and inference,” Journal of the Royal Statistical Society: Series A (Statistics in Society), vol. 176, no. 3, pp. 703–722, 2013.

[2] D. Thissen, L. Steinberg, and H. Wainer, “Detection of differential item functioning using the parameters of item response models.,” 1993.

[3] P. W. Holland and H. Wainer, Differential item functioning. Routledge, 2012.

[4] G. King, C. J. Murray, J. A. Salomon, and A. Tandon, “Enhancing the validity and cross-cultural comparability of measurement in survey research,” American political science review, vol. 98, no. 01, pp. 191–207, 2004.

[5] M. G. Kendall, “A new measure of rank correlation,” Biometrika, pp. 81–93, 1938.

[6] H. Valizadegan, R. Jin, R. Zhang, and J. Mao, “Learning to rank by optimizing ndcg measure,” in Advances in neural information processing systems, 2009, pp. 1883–1891.

[7] A. R. Daykin and P. G. Moffatt, “Analyzing ordered responses: A review of the ordered probit model,” Understanding Statistics: Statistical Issues in Psychology, Education, and the Social Sciences, vol. 1, no. 3, pp. 157–166, 2002.

[8] R. Yan and J. Hegeman, Using Polling Results as Discrete Metrics For Content Quality Prediction Model. Google Patents, 2014.

[9] R. D. McKelvey and W. Zavoina, “A statistical model for the analysis of ordinal level dependent variables,” Journal of mathematical sociology, vol. 4, no. 1, pp. 103–120, 1975.

[10] S. Berry, J. Levinsohn, and A. Pakes, “Automobile prices in market equilibrium,” Econometrica: Journal of the Econometric Society, pp. 841–890, 1995.

[11] U. Rao, T. Sidhartha, K. R. Harker, A. S. Bidesi, L.-A. Chen, and M. Ernst, “Relationship between adolescent risk preferences on a laboratory task and behavioral measures of risk-taking,” Journal of Adolescent Health, vol. 48, no. 2, pp. 151–158, 2011.

[12] M. E. Ben-Akiva and S. R. Lerman, Discrete choice analysis: theory and application to travel demand, vol. 9. MIT press, 1985.

[13] R. D. Luce, Individual choice behavior: A theoretical analysis. Courier Dover Publications, 2005.

[14] L. Muchnik, S. Aral, and S. J. Taylor, “Social influence bias: A randomized experiment,” Science, vol. 341, no. 6146, pp. 647–651, 2013.

[15] M. J. Salganik, P. S. Dodds, and D. J. Watts, “Experimental study of inequality and unpredictability in an artificial cultural market,” science, vol. 311, no. 5762, pp. 854–856, 2006.

[16] J. B. Schafer, D. Frankowski, J. Herlocker, and S. Sen, “Collaborative filtering recommender systems,” in The adaptive web, Springer, 2007, pp. 291–324.

[17] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, “Item-based collaborative filtering recommendation algorithms,” in Proceedings of the 10th international conference on World Wide Web, 2001, pp. 285–295.

[18] P. Resnick and H. R. Varian, “Recommender systems,” Communications of the ACM, vol. 40, no. 3, pp. 56–58, 1997.

[19] Q. V. Le, “Building high-level features using large scale unsupervised learning,” in Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on, 2013, pp. 8595–8598.

[20] Glimcher, P.W. Decisions, uncertainty, and the brain: The science of neuroeconomics. MIT press, 2004.

[21] Kahneman, D. "Maps of bounded rationality: Psychology for behavioral economics." *American economic review* (2003): 1449-1475.

[22] Peysakhovich, A. and Karmarkar, U. “Asymmetric Impacts of Favorable and Unfavorable Information on Decisions Under Ambiguity,” *mimeo.*

[23] Fudenberg, D. and Peysakhovich, A. “Recency, Records and Recaps: Learning and non-equilibrium behavior in a simple decision problem,” *Proceedings of the ACM Conference on Economics and Computation 2015 (EC14)*

[24] Junfeng Pan, He Xinran, Ou Jin, Tianbing XU, Bo Liu, Tao Xu, Yanxin Shi, Antoine Atallah, Ralf Herbrich, Stuart Bowers, Joaquin Quiñonero Candela, “Practical Lessons from Predicting Clicks on Ads on Facebook,” *International Workshop of Data Mining for Online Advertising (ADKDD)*