The “Expression Gap”: Do you Like what you Share?

Atish Das Sarma
eBay Research Labs
atish.dassarma@gmail.com

Si Si
University of Texas-Austin
ssi@cs.utexas.edu

Elizabeth F. Churchill
eBay Research Labs
churchill@acm.org

Neel Sundaresan
eBay Research Labs
nsundaresan@ebay.com

ABSTRACT

While recommendation profiles increasingly leverage social actions such as “shares”, the predictive significance of such actions is unclear. To what extent do public shares correlate with other online behaviors such as searches, views and purchases? Based on an analysis of 950,000 users’ behavioral, transactional, and social sharing data on a global online commerce platform, we show that social “shares”, or publicly posted expressions of interest *do not* correlate with non-public behaviors such as views and purchases. A key takeaway is that there is a “gap” between public and non-public actions online, suggesting that marketers and advertisers need to be cautious in their estimation of the significance of social sharing.

1. INTRODUCTION

In recent years, significant effort has been expended trying to capture and understand online user behavior as a basis for tailoring users’ experiences. Contemporary product and service marketing through personalized targeting makes the assumption that presentation of similar or related content will result in greater likelihood of repeat or further consumption. Recommendations are selected based on users’ prior behaviors and expressed preferences.

Increasingly “social signals” such as “likes” and “shares” are also considered to be a reflection of a user’s interests. Commerce platforms also offer “share widgets” for easy posting to social networks and our results are based on such data gathered from eBay.

For promoters of events, products, and services there are two key questions regarding such social sharing: (1) does exposure to new potential consumers increase the number of people consuming the service or product, and (2) do these socially expressed interests actually correlate with a users’ intent to purchase and with their own consumption and purchasing decisions. While much research has been conducted on the first question, e.g., [2], little research has directly addressed whether social shares from commerce platforms reflect a stronger likelihood of actual purchase by the individual who posts them. In this paper we address this second question.

Our results indicate that, at the aggregate level, socially expressed interests do not correlate strongly with intrinsic tastes. We show that consumption (measured as either purchases or browsing behavior) often diverges from expression (measured as shares to external social networks). We establish that this discrepancy is beyond random behavior and call this the *expression gap* (see Figure 1). We further explore the expression gap when various signals for consumption are combined together, specifically a linear combination of purchasing and browsing behavior. Our analysis shows that this explains only a part of the behavioral gap.

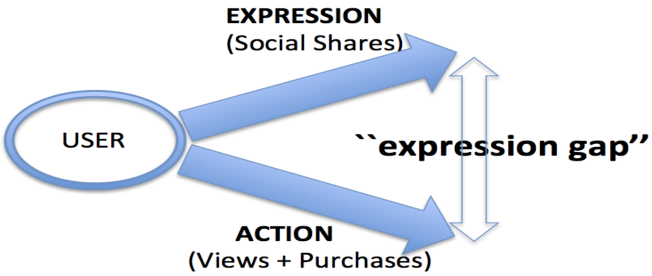


Figure 1: Diagram showing the notion of “expression gap”.

Related Work. Crandall et al. [4] show that similarity between users can predict social interactions over long periods of time. Also [5] consider homophily in social networks and conclude that modest preferences to similar others can be amplified by biased selection. [3] undertake a game theoretic exploration of how opinions evolve in the presence of friendship, and [1] study the impact of social influence on behavior. Several recent papers [7, 6] have studied incorporating social interaction for e-commerce recommender systems.

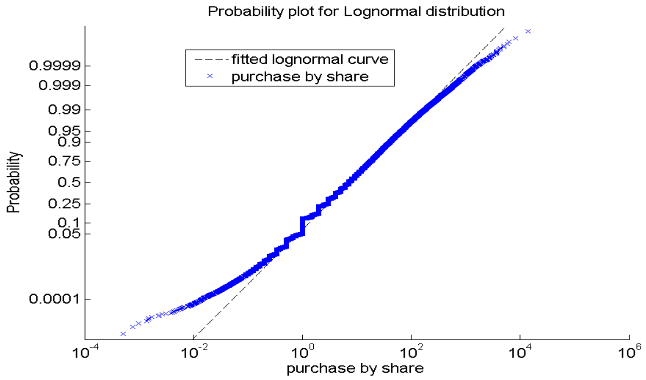
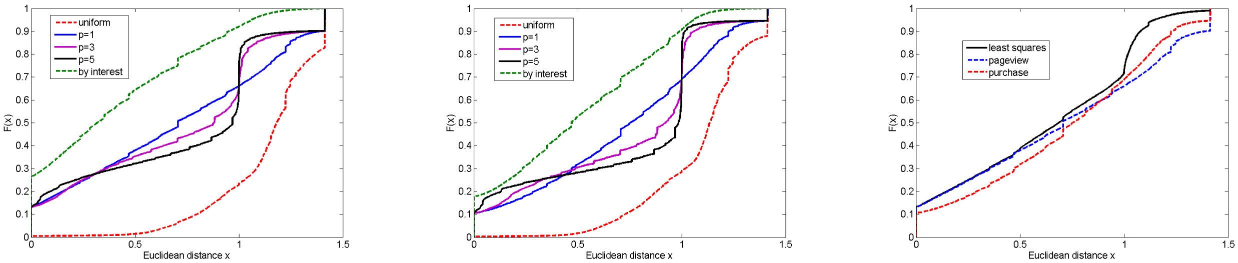


Figure 2: CDF plot: x = #purchases/#shares and y = fraction of users with that ratio. Dashed line indicates log-normal fit.

WWW’14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04
http://dx.doi.org/10.1145/2567948.2577295.



(a) page view-share gap (b) purchase-share gap (c) linear combination

Figure 3: CDF plots of expression gap for shares vs. (a) page views, (b) purchase, (c) linear combination of page views and purchases.

2. DATASET AND STATISTICS

We gathered data from eBay and focus only on buyers as sellers may be motivated to share/promote items they are selling. Our dataset included almost a million unique users (active between Dec 2012 and June 2013), several million shares, tens of millions of purchases, and over 60 million page views. The marketplace includes 39 self-explanatory top-level product type categories such as Fashion, Electronics, Real Estate, Home & Garden, and Sports.

The CDF plot in Figure 2 indicates that there are several users who buy a lot but almost never share, and vice versa, as is to be expected. This differing behavior, however, is expected. The next section explores this at a content category level, establishing a gap even among users that share and consume a lot.

3. MEASUREMENTS AND RESULTS

We first introduce some notation. We define a sharer as a user who has shared at least one item. For each sharer i, we have the following. e_i is a vector that distributes user i's shares into the 39 top-level categories, normalized to form a probability distribution. So $e_i(j)$, shows the fraction of shares made by i that fall in the j^{th} category. Similarly, based on user i's page view and purchase history, we define $c_{i,1}$ or $c_{i,2}$. In order to investigate whether a user's shares are random behavior or closely aligned with personal interests, we estimate the distances between two kinds of consumption c_i (either $c_{i,1}$ or $c_{i,2}$) and expression (e_i) using Euclidean distance. Further, by e_i^p, we denote the vector obtained by powering each value in e_i to the p-th power, and then normalizing it back to a probability vector (similarly for $c_{i,1}^p$ or $c_{i,2}^p$). This skews vectors in favor of the *stronger* dimensions.

Random processes for comparison. If we simply compute $d(e_i, c_{i,1})$ and $d(e_i, c_{i,2})$, the distance values would not mean much in isolation, since it is hard to interpret whether a specific value is "similar" or "different". We therefore use two baselines. The first is obtained by generating a share vector $\bar{e}_i$ uniformly at random across the dimensions (respecting the number of shares for each user).

The second baseline mimics a situation of consumption and expression arising from the same *distribution*. So the real share vector e_i is compared against a simulated "interest-based" share vector $\tilde{e}_i$. $\tilde{e}_i$ is generated by sampling shares for each user, from its own page view or purchase distribution (as the case may be). $d(\tilde{e}_i, c_{i,1})$ and $d(\tilde{e}_i, c_{i,2})$ are the resulting baselines (notice this is a significantly stronger test than comparing against the baseline of zero gap).

Do shares correlate with consumption? We show the main results in Figure 3. In Figures 3a and 3b, we show five lines, two corresponding to the two baselines based on uniform and interest-based sampling (shown in dashed lines), and three lines each based on the distances $d(e_i^p, c_{i,1}^p)$ and $d(e_i^p, c_{i,1}^p)$, for $p = 1, 3, 5$. Since these are CDF plots, higher lines reflect smaller distances overall. We see that increasing p does not help in reducing the overall gap. The red dashed line corresponds to distances if the shares were performed uniformly at random, $d(\bar{e}_i, c_{i,1})$ or $d(\bar{e}_i, c_{i,2})$. Further, the green dashed line corresponds to the distances with the simulated share, i.e. if the shares were performed according to the page view or purchase decisions, $d(\tilde{e}_i, c_{i,1})$ or $d(\tilde{e}_i, c_{i,2})$. The green dashed line stochastically dominates all the solid lines. This suggests that shares and actual purchases tend to be significantly different even though users clearly make deliberate decisions when sharing.

Test through linear combinations. Figure 3c discusses the setting when we optimize a parameter independently for each user to minimize the gap between their expression and consumption. In particular, we compute for each user i a parameter α_i to minimize $d(e_i, \alpha_i \cdot c_{i,1} + (1 - \alpha_i) \cdot c_{i,2})$. The solid black line corresponds to the resulting distance distribution, while the red and blue dashed lines correspond to the previous distributions of $d(e_i, c_{i,1})$ and $d(e_i, c_{i,2})$ respectively. We omit plotting the distribution of the values of α_i due to space constraints; it was noticed that around 30-40% users each had an optimal $\alpha_i = 0$ or $\alpha_i = 1$, while the remaining α_i values were uniformly spread across $(0, 1)$. We see that the black solid line only results in a marginal *lift* in stochastic domination. This is somewhat surprising, suggesting that there remains a significant component of expression gap that remains unexplained even as differing signals of consumption are combined together.

This gap across the figures illustrates that the strongest signal of true deep-seated interests are fundamentally different and do not necessarily manifest in the form of socially expressed preferences.

4. REFERENCES

[1] A. Anagnostopoulos, R. Kumar, and M. Mahdian. Influence and correlation in social networks. In *KDD '08*, pages 7–15, 2008.

[2] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts. Everyone's an influencer: quantifying influence on twitter. In *WSDM*, 2011.

[3] K. Bhawalkar, S. Gollapudi, and K. Munagala. Coevolutionary opinion formation games. In *STOC '13*, pages 41–50, 2013.

[4] D. J. Crandall, D. Cosley, D. P. Huttenlocher, J. M. Kleinberg, and S. Suri. Feedback effects between similarity and social influence in online communities. In *KDD*, pages 160–168, 2008.

[5] G. Kossinets and D. J. Watts. Origins of homophily in an evolving social network1. *American Journal of Sociology*, 115(2), 2009.

[6] C. Lam. Snack: incorporating social network information in automated collaborative filtering. In *EC*, pages 254–255, 2004.

[7] J. Leskovec, L. A. Adamic, and B. A. Huberman. The dynamics of viral marketing. *ACM Trans. Web*, 1(1), 2007.