Proceedings of the Tenth International AAAI Conference on
Web and Social Media (ICWSM 2016)

Measuring the Efficiency of Charitable Giving with Content Analysis and Crowdsourcing

Ceren Budak
University of Michigan
cbudak@umich.edu

Justin M. Rao
Microsoft Research
justin.rao@microsoft.com

Abstract

In the U.S., individuals give more than 200 billion dollars to over 50 thousand charities each year, yet how people make these choices is not well understood. In this study, we use data from CharityNavigator.org and web browsing data from Bing toolbar to understand charitable giving choices. Our main goal is to use data on charities' overhead expenses to better understand efficiency in the charity marketplace. A preliminary analysis indicates that the average donor is "wasting" more than 15% of their contribution by opting for poorly run organizations as opposed to higher rated charities in the same Charity Navigator categorical group. However, charities within these groups may not represent good substitutes for each other. We use text analysis to identify substitutes for charities based on their stated missions and validate these substitutes with crowd-sourced labels. Using these similarity scores, we simulate market outcomes using web browsing and revenue data. With more realistic similarity requirements, the estimated loss drops by 75%—much of what looked like inefficient giving can be explained by crowd-validated similarity requirements that are not fulfilled by most charities within the same category. A choice experiment helps us further investigate the extent to which a recommendation system could impact the market. The results indicate that money could be re-directed away from the long-tail of inefficient organizations. If widely adopted, the savings would be in the billions of dollars, highlighting the role the web could have in shaping this important market.

Introduction

In the U.S., individuals annually give more than 200 billion dollars to over 50 thousand charities. Despite the large economic consequences of these decisions, a review of the literature reveals that basic questions, such as how people decide how much money to donate or which charities to support, are not well understood (Andreoni 2006). This stands in contrast to our understanding of choices for traditional consumer goods, which is one of the primary areas of study in economics and marketing. One reason for the knowledge disparity is the ease with which consumer goods can be quantified by their attributes. For example, in a "demand system" for televisions, an analyst or researcher can easily represent products as feature vectors that include size, resolution, manufacturer, price, etc. Television producers compete in this market by making advances in features or cutting price, and this can be quantitatively studied with well-developed statistical techniques. In contrast, charities have proved harder to distill to key features that drive consumer choice. Put another way, if we view charities as "products," then it is fair to say that economic analysis has not used "product features" with comparable quantitative rigor as traditional markets (Auten, Sieg, and Clotfelter 2002; Andreoni 2006).

In this paper we use a combination of text analysis, web activity logs and crowdsourcing to help bridge this gap. The central question we address is how much money is "wasted" through donations to inefficiently run charities for which there is a close substitute. Charities, like firms, vary quite widely in the quality of their "products." The difference is that when a consumer buys a traditional good she gets feedback on product quality. Since this feedback is actively shared via reviews, word-of-mouth and online "star ratings," firms that produce goods of low quality relative to similarly priced competitors find it hard to stay in business. No such feedback loop exists for charities. A donor writes a check or makes an online contribution and rarely gets to observe the provision of charitable services (Weisbrod and Dominguez 1986). While ratings agencies like Charity Navigator attempt to fill the void, it has been previously documented (Hager and Greenlee 2004) that charities exhibit considerable differences in their ability (or willingness) to put donations towards the charitable mission as opposed to administrative costs, staff salary, fundraising events, etc. At first glance, this seems to be strong evidence the market is inefficient—if people just made more informed choices, then "bad charities" would be forced out of the market.

An obvious flaw in this line of reasoning is that it presumes that there are good substitutes for the offending charities. Someone's choice to donate to a given charity presumably depends on factors such as the charitable mission, the location, religious affiliation and so forth. It is easy to imagine, then, that there are many niches in this space and people simply prefer to have a close match to their preferences at the expense of the efficiency in which the donation goes to the cause in question.

To investigate the extent to which this argument is supported by data we first quantify the similarity between char-

ities based on their stated missions. In the U.S., charities are required to state their mission and charitable works on their IRS 990 forms to qualify for tax exempt status. These approximately 100 word documents are ideal textual input for the task at hand. We scrape these descriptions from the charity ratings website Charity Navigator. We also collect additional information reported by the site: two-level categories (e.g. Health-Medical Research) and efficiency score (a well-respected measure for what fraction of each dollar donated that goes to the charitable mission), the total annual contributions, and the location information. We complement the data on total contributions with site visits to donation pages as measured through the Bing toolbar over a 17 month period. These data capture a broader measure of charity support, as they cannot be swayed by a single large donor. Overall, the data provide detailed information on 7,869 charities, that collectively account for more than half of the charitable donations in the U.S. ($\approx$ 116 billion dollars per year).

We start by looking at the efficiency distribution within Charity Navigator's two-level categories, where we normalize efficiency relative to the best charity within the category. If this comparison charity is an acceptable substitute, then this distribution has the interpretation of "efficiency loss." Using this categorization produces quite dispersed loss—more than half donations and online visits goes to charities with greater than 15% loss.

These relatively broad categories, however, may fail to capture important aspects of the charitable mission. For example, someone interested in donating to a HIV charity may find a cancer charity a poor substitute. We measure the similarity between any two charities by taking the tf-idf weighted cosine distance of the mission descriptions. To calibrate and validate this measure, we turn to crowdsourcing labels on Amazon Mechanical Turk. The results reveal that cosine distance is a strong predictor of similarity.

Using these scores and information on the location where charitable services are administered, we compute the loss distributions with increasingly strict requirements of similarity. These simulations provide informative bounds for the efficiency of the marketplace. For each charity this process specifies a similarity requirement that determines what charities form the "substitution set." We also examine the requirement that a comparison charity operates in the same location, as the web browsing logs reveal a strong "home bias," indicating that people would hold this as an important factor. For a threshold of 0.2, a value that ensures most users find the charities to be rather similar, we still observe substantial inefficiency. However, when the location requirement is added, this loss falls by about 75%. Still, there is a meaningful long-tail of inefficient organization that have close substitutes. Collectively in our sample this amounts to over 4 billion dollars in contributions. We further show that the offending organizations tend to be larger, which is intuitive as a large charity is more likely than a niche organization to have an acceptable, well-run substitute. At very high levels of required similarity the total loss falls further still, but a long-tail of inefficient laggards remain.

One way that technology could improve marketplace efficiency is a recommendation system similar to those used on e-commerce and streaming media platforms. For example, before "check out" a list of comparison charities could be given, along with their efficiency scores and other relevant characteristics that help identify "good matches." The degree to which such a technology could impact the market is hard to pin down precisely, but our initial results indicate there is substantial room for efficiency gains. To further refine our estimate we run an Amazon Mechanical Turk experiment to estimate how consumers would respond to a recommendation system using hypothetical questions. Subjects were told to imagine they had decided to give to a focal charity with 70% efficiency and were presented with real alternatives that were specified to have either 80% or 90% efficiency and had varying similarity. Subjects were in general willing to switch to the more efficient charities and this willingness increased with similarity, as measured by cosine distance, and the efficiency gain. While one cannot derive externally valid point estimates from such hypothetical questions, we believe the directional results are on firm ground and that the results point to the promise of such a recommendation system. We leave a more rigorous study of such systems to future work.

Overall the results indicate that the marketplace for charitable giving is not as inefficient as it seems at first glance. Much of the putative loss can be attributed to a lack of close substitutes. However, even when a high standard of similarity and location-matching is required, a sizable inefficiency remains. And while it is not known the degree to which consumers would opt for more efficient choices if provided with a "frictionless" user interface, our preliminary analysis indicates there is much promise in this approach.

Related Work

Previous work has investigated the extent to which "price" impacts charitable giving (Randolph 1995; Auten, Clotfelter, and Schmalbeck 2000; Auten, Sieg, and Clotfelter 2002; Bakija, Gale, and Slemrod 2003; Karlan and List 2007). The two primary drivers of the effective price are the marginal tax rate and the efficiency of the charity (which governs how much of donations is lost before being administered to the charitable mission). Responses to price have also been evidenced by a higher propensity to give when the gift if matched by a third party (Karlan and List 2007; Eckel and Grossman 2008). To understand the role of taxes, suppose that there is a flat income tax that is raised from 20% to 30%. If donations are deductible, then this amounts to a 10% reduction—donating $100 now effectively costs $70 in post-tax income instead of $80. A central question is whether demand is "elastic"—does a 1% price drop lead to a greater than 1% increase in donations? Unfortunately, due in part to the challenges mentioned in the introduction, the literature offers discordant answers—the two most widely cited papers (Randolph 1995; Auten, Sieg, and Clotfelter 2002), which use very similar data of individual income returns from the IRS, report estimates that straddle this cutoff by a wide margin.

The impact of charity efficiency on donor behavior has received less attention. Surveys (Glaser 1994; Bennett and Savani 2003) and small scale experiments (Parsons 2007; Buchheit and Parsons 2006) suggest that donors care about

the efficiency of charitable organizations they contribute to. (Gordon, Knock, and Neely 2009) uses a panel of charities from Charity Navigator and finds that increases in the "star rating" over time are correlated with increases in contributions, controlling for other factors. This is evidence that consumers value higher rated charities. The point estimates of the dollar value of increased donations are modest, however, indicating that most contributions are not driven by the ratings change. We note that none of these studies take substitutes into consideration and characterize charitable giving choices in the presence of acceptable substitutes. Recent work (Karlan and Wood 2014) has found that providing scientific evidence on the effectiveness of a charity's interventions (e.g. reduction in malaria from mosquito nets in an AB test) can either increase or decrease giving, depending on donor "type." Finally, charities have been shown to use accounting practices that make it more likely that they will get a high rating from organizations like Charity Navigator, which indicates a belief on their end that these ratings matter in the minds of donors (Hager and Greenlee 2004).

Data

In characterizing charitable giving behavior we make use of the following data sets:

Charity Navigator Data: Charity Navigator (CN) (www.charitynavigator.org) is an independent non-profit that has assessed over 8,000 401c3 organizations in the United States based on organizational efficiency and capacity. We scrape CN to collect data about all such charitable organizations. We restrict the charities of relevance to those that are in good standing which reduces the number to 7869 charities which collectively received $\approx$ 116 Billion Dollar contributions over the past year. Given this data set, we identify the following charity features:

1. **Title and Webpage:** Charity name and homepage
2. **Overall Contributions:** Total charitable contributions listed in IRS 990 Forms.
3. **Overhead:** There are three high level expenses that help define the efficiency of a charitable organization: *fundraising, administrative and program expenses*. We define $\frac{fundraising+administrative}{fundraising+adminstrative+program}$ expenses as the *overhead* of a charity. The lower this overhead is, the more of the donated money is spent directly on the cause of the organization.
4. **Category:** CN tags each charity with a first level (e.g. Health), and a second level category (e.g. Health: Medical Research). CN has 11 first and 35 second level categories.
5. **Mission statement:** Mission statements as listed by CN.
6. **Location:** Headquarters of the charity listed on CN.
7. **Locational Focus:** In addition to the charity address, we identify the city, state, and country mentions in charity mission statements to identify their locational focus. This location might be different from the location of the charity (e.g. internationally focused charities).

Charity Websites Scraped From the Web: CN lists the homepage for each organization it rates. We use this data to identify the donation traffic each charity receives. However, homepage traffic is not necessarily indicative of intended donations. For instance, a large number of people visit plannedparenthood.org to seek health related knowledge with no intention of making a monetary donation. Thus, we develop a web scraper script that identified donation related links on charity homepages. This resulted in a list of links that are highly indicative of charitable giving. This list includes 30911 links from 7869 charity domains.

Bing toolbar dataset: To identify online donation traffic each charitable organization receives, we first examined the complete web browsing records for U.S.-located users who installed the Bing Toolbar, an optional add-on for the Internet Explorer web browser. We detect charitable giving attempts by identifying visits to donation links described above. For each such visit, we record the timestamp, the deidentified toolbar user id and the zipcode the visit was initiated from. We collected data from 25-November-2013 to 15-May-2015. This provides charity visits of $\approx$3.6 million from $\approx$1.3 million unique web users. The time frame covered by the Bing toolbar dataset also roughly corresponds to the time period IRS 990 forms cover[1].

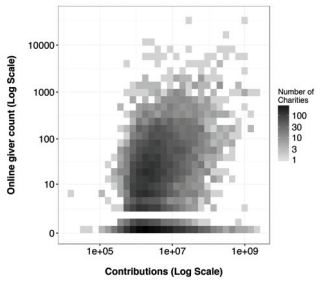
Related work has typically either used individual level tax return data (e.g. (Auten, Sieg, and Clotfelter 2002; Randolph 1995)), or aggregate contribution data (e.g. (Gordon, Knock, and Neely 2009; Andreoni and Payne 2003)). In this paper the Bing Toolbar and IRS 990 forms provide data on both of these pieces.[2] We refer to these two data sets as "online donor" and "overall contributions."

In Figure 1(a) we show the relationship between online donors and overall contributions. These measures are positively correlated but the correlation is modest ($\approx$0.2). This result, while interesting, is not unexpected given the importance of very large donors in fundraising (Andreoni 2006). Figure 1(b) provides another perspective, showing how the relationship between online donors and overall contributions varies across different charity categories. The x-axis is normalized such that the category with the lowest contributions to online ratio (Animals) is assigned a value of one—other categories can be evaluated in relation to this category. Increasing values indicate more dollars in contributions relative the number of online donors. The results show wide difference across categories. Education charities have the lowest online presence relative to overall contributions, a 33:1 difference as compared to animal related charities.

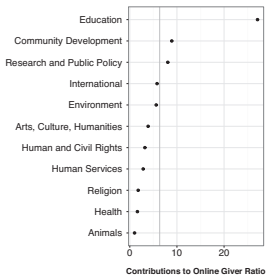
We further investigate these differences by looking at how unigrams in charity mission statements predict total contributions and the number of online donors. We do so with a linear regression model that is penalized with the L2-norm

[1] Data was collected from CharityNavigator on March 10, 2015.

[2] There are three main distinctions between the Bing Toolbar and the IRS forms. First, Bing Toolbar data includes intended donations while IRS forms capture only transacted donations. Second, toolbar data identifies number of donors while IRS forms list dollars contributed. This distinction is particularly important since most charitable contributions are dominated by large amounts of money contributed by big donors (Clotfelter 2001; Andreoni 2006). Finally, toolbar data is restricted to online users while the donations listed on IRS forms are online and offline.



(a) Scatter Plot



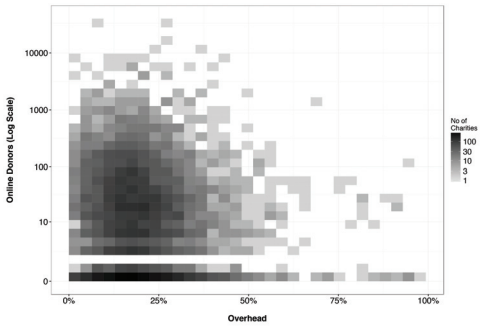
(b) Contributions/Online Ratio Across Categories

Figure 1: Relationship between online presence and overall dollar contributions

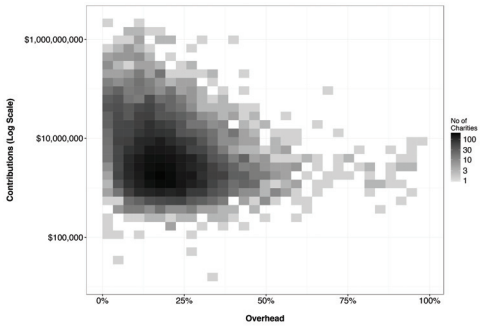
Online Donors		Overall Contributions	
Popular	Unpopular	Popular	Unpopular
nationwide	actively	undergraduate	choices
undergraduate	rning	million	often
metropolitan	systems	millions	religious
race	workshops	offices	connect
cure	business	university	workshops
wishes	providers	institution	understand
charity	economic	philanthropy	almost
million	benefit	faculty	group
dogs	focused	liberal	disabled
scientific	different	donors	spay
television	leaders	outstanding	using
sufficiency	currently	fight	professionals
bitat	reproductive	highest	outdoor
name	council	products	equip
affected	series	distribution	relationship
military	institute	largest	promoting
animals	legal	ideas	board
continuing	citizens	philanthropic	groups
colleges	using	graduate	information
news	sense	relief	empowering

Table 1: Most Predictive Words To Identify Popular and Unpopular Charities Online and Offline

where the LHS is defined as 1) log of number of online donors an organization has 2) log of overall contributions an organization receives. We identify the features (words) that have the highest and lowest weights which provide the list of most predictive words for popular and unpopular charities respectively. The results are given in Table 1. Online donors vs. overall contributions pull out largely different features. While some of the differences are categorical ("dog" vs. "university"), there are features that capture a more subtle difference as well. For charities with high overall contributions, it is picking up words associated with large organizations (e.g. "institution" and "million") whereas for online donors it picks more "aspirational" words (e.g. "race", "cure" and "wishes"). This exploratory analysis reveals how different aspects of charitable organizations predict attention from different segments of the donor base.



(a) Online Donors



(b) Overall Contributions

Figure 2: Relationship between a) online donors, b) overall contributions and charity overhead

Preliminary Analysis

Our goal in this paper is to quantify the inefficiencies in the charity marketplace. To do so, we rely on the overhead measure and start by asking the following question: *How well do charities with high/low expense overheads perform in terms of overall contributions and online traffic?* Figure 2 answers this question. Really inefficient charities (with $>$50% overhead) do poorly both in terms of overall contributions and online traffic. However, the discriminative power of this feature is less clear beyond this point. This indicates that there is potentially a large "waste" in the charitable marketplace. Why does that "waste" exist? Is it real? We address these questions in the remainder of the paper.

Our first step is to use browsing behavior to infer preferences over charity category and the location. Both analyses suggest that donors appear to have preferences that limit the "acceptable options," potentially to those that have relatively large overhead.

Home bias: "Home bias" is a term used to refer to preferences of individuals for organizations located near them. While it is often called a "bias," a more neutral interpretation is that this is a feature that impacts utility. We investigate the role of location using the web browsing data set. In 39% of online giving cases, donors choose to contribute to charities in their home state. In 13% of cases, donors give to charities in their city. These figures far exceed what one would expect due to chance alone. In Figure 3, we demonstrate how home bias varies across charity types. The x-axis provides the fraction of donations that are made to a charity in the same state

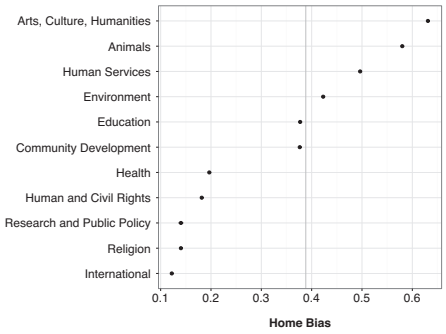


Figure 3: Tendency of donors to donate to charities in their own state

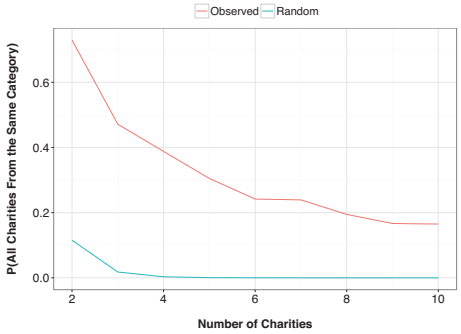


Figure 4: The likelihood that a user donating to k charities makes donations to charities of a single category

as the user. The y-axis is ordered from most to least local. We also plot standard errors but given the large amount of data, they are near zero and are not visible in the plot.

Overall, the figure shows that donors tend to make more local decisions for arts, animals and human services related charities. An opposite pattern is observed for international, religious and research related charities. A very similar pattern is observed for in-same-city analysis. These results reveal that locational requirements, in practice, are more important for certain categories.

Categorical choices: Do people focus their donations on charities from one or few categories or do they spread their donations more or less evenly across charities from different categories? To answer this question, we consider toolbar users who showed intend to donate to at least 2 charities. We identify the fraction of web users with an intend to donate to $k \geq 2$ where all k charities are from the same category and compare that to the equivalent of this number where users choose charities from the list of ≈8,000 charities at random. The results presented in Figure 4 indicate that donation behavior is far more concentrated in a single category than we would expect by chance, confirming our prior that people have stable preferences over high-level charitable missions.

Based on these preliminary analysis, we conclude that donor preferences over location and mission of organizations might result in a small number of effective options for donors, which can help explain the relatively large number of charities that have large overheads. Next, we define various choice models based on these findings and estimate the inefficiencies in the current charity marketplace accordingly.

Defining Choice Models

Here we define various choice models motivated by the evidence presented in the previous section:

1. First-Level Category Model: Under this simplistic choice model, we assume that charities under the same first-level CN category (e.g. Health) constitute a valid substitute for one another.

2. Second-Level Category Model: Under this choice model, we assume that charities under the same second-level CN category (e.g. Health: Medical Research) constitute a valid substitute for one another.

3. Second-Level Category + Location Model: Under this choice model, we assume that charities that are under the same second-level CN category and conduct charitable works in the same location constitute a valid substitute for one another[3].

4. Similarity=k Model: While analysis in the previous section suggests the CN categories captures some aspects of donor preferences, they might not provide a fine-enough level of information to truly identify the choices for at least a segment of donors. For instance, a donor interested in making a contribution to a cancer research charity might find another health related charity a poor substitute, no matter how much more efficient the other charity is. Here, we move beyond the categories given by CN and provide a measure of similarity between any two charities using their mission statements. This model is described in detail below.

5. Similarity=k + Location Model: Under this choice model, we assume that charities that have a similarity measure of at least k and are focused on the same location constitute a valid substitute for one another.

6. Data-driven Model: We estimate the importance of charity similarity, location, and overhead difference in determining whether a particular charity is a good substitute for another through an Amazon Mechanical Turk experiment. This model and the experiment are described in detail below.

Similarity=k Model

We use a cosine similarity measure with tf-idf weighting that is commonly used in information retrieval (Baeza-Yates, Ribeiro-Neto, and others 1999).[4] This straightforward approach proved successful in identifying similarity between charities as demonstrated by our Amazon Mechanical Turk experiment below. In the future, we aim to investigate the use of other similarity measures to build towards a charitable giving recommendation system.

[3]Charities that do not have a particular geographical focus are modeled as having a worldwide focus and therefore can only be matched to other worldwide focused charities.

[4]We remove location mentions from the mission statements when computing this measure as our goal is to determine cause relevancy as opposed to location. Note that this will be combined with location information to form a more strict model later.

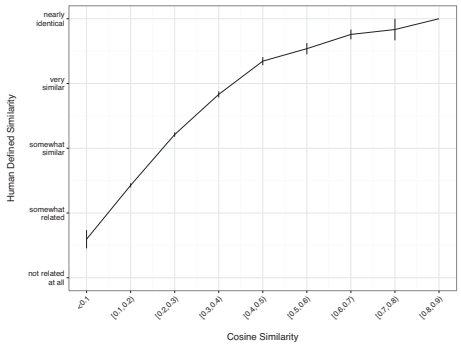


Figure 5: Relationship between cosine distance between two charities and worker evaluation of their similarity.

Worker qualifications. To ensure high-quality ratings, we required that workers: (1) reside in the U.S.; (2) had successfully completed at least 1000 Mechanical Turk "Human Intelligence Tasks" (HITs); and (3) had an approval rate of at least 98%. In order to ensure a large, representative labor force, workers were restricted to at most 30 HITs.

Worker instructions. Workers were given the description of a focal charity and asked how similar 5 comparison charities were to this focal charity (not related at all, somewhat related, somewhat similar, very similar, nearly identical).

Data Sampling. We choose 330 focal charities at random stratified over the 11 first-level categories. For each focal charity, we choose 1 charity at random from the same second-level CN category and 4 charities according to their cosine similarities to the focal charity. This allows comparison between CN category and cosine similarity matches.

Evaluation. Figure 5 gives the relationship between similarity score and mean worker evaluation. Below 0.1, two charities were judged to be unrelated. At 0.2, the mean judgment is close to "somewhat similar" and increases to "very similar" at around 0.3-0.4. Above this value the ratings converge to "nearly identical." These results validate our similarity score and provide useful calibration for the similarity-k models used in the next section. In comparison, the average worker evaluation for comparison charities from the same second level category is 1.6, which corresponds to a value between "somewhat related" to "somewhat similar".

Considering all charities in our dataset: if we take two charities in the CN top-level, the average cosine distance is 0.09, roughly corresponding to "somewhat related". The similarity average for two-level category is 0.11, corresponding to"somewhat related" to "somewhat similar" (in agreement with the Mechanical Turk findings).

Inter-rater reliability. Each comparison choice task was given to 3 people. In total, 111 workers participated in this study. We compute inter-rater reliability in two ways. First, we require that at least two of the human judges choose the exact same value in the 5-point scale. Second, we require judgments to differ by at most a total of 1 point on the scale (e.g. 2 judgments at very similar, 1 at somewhat similar). IRR is 0.74 and 0.57 under these definitions respectively. Overall, this shows high inter-rater reliability.

Data-driven Model

In the previous section, we demonstrated that cosine similarity, and to a lesser extend CN categories, can be used to identify similar charities for a given focal charity. Here we ask the next natural question: *"Can one alter the charitable giving decisions of a user by providing information on the efficiency of a charitable organization?"* If so, how does the expense overhead, location, and similarity of the substitution affect those choices? Definitively answering these complex questions would require substantial real-world experimentation—here we conduct a hypothetical choice study to provide some initial insights and help calibrate our choice simulations in the next section.

Worker qualifications. We use the same qualifications as the *similarity=k* experiment described before.

Worker instructions. Workers were again given a focal charity. This time they were told to imagine they believed in the cause and were about to donate to this charity. They were further told that 70 cents for every dollar donated would go towards the charitable mission described (30% overhead). They were then presented with 5 comparison charities with potentially different missions. In the first condition, subjects were informed these charities passed on 80 cents on the dollar to charitable works. In the second, the efficiency was increased to 90%. Subjects were asked how likely they would be to switch to the comparison charities (5 point scale from "definitely not" to "definitely").

Data Sampling. We chose 440 focal charities at random stratified over the 11 first-level categories. For each focal charity, we identified 12 possible candidates to present to the workers: *1) 1 first-level category match, 2) 1 second level category match, 3) 1 first level category + location match, 4) 1 second level category + location match, 5) 4 cosine similarity matches of varying levels, and 6) 4 cosine similarity matches of varying levels + location match*. Five options were chosen at random among these 12 candidates. Each comparison was performed by 3 workers. Our experiment involved 232 unique participants.

Results. In total, 232 workers participated in this study. Table 2 summarizes the results with a regression where the dependent variable is the switching score (1–5, 5 indicates the participant would "definitely" switch). The results can be summarized as follows:

1. We see that the baseline tendency to switch (switching from the focal charity to a charity that has 0 cosine similarity to the focal charity and spends 80% of charitable donations on the program) is 1.75 which is in between unlikely (=1) and somewhat likely (=2).
2. The alternative charity being in the same high level category increases this value by 0.09 points, as denoted in the Type1 row. The effect is slightly higher when the alternative charity is in the same second level category (0.1). Both of these values are significant at the 0.05 level.
3. Higher cosine similarity results in higher likelihood of switching. In particular, the cosine similarity moving from 0 to 1 increases the switching choice by 1 point in the 5 point scale. We can also observe that a cosine similarity of 0.1 results in approximately the same lift as Type

	Estimate	Std. Error
(Intercept)	1.75^{***}	(0.04)
LocationsMatch	0.03	(0.02)
EfficiencyLevel	0.18^{***}	(0.02)
Type1	0.09^{*}	(0.04)
Type2	0.10^{*}	(0.04)
TDIDF	1.00^{***}	(0.13)
AIC	43441.24	
BIC	43493.66	
Log Likelihood	-21713.62	
Deviance	20744.53	
Num. obs.	13200	

$^{***}p < 0.001, ^{**}p < 0.01, ^{*}p < 0.05$

Table 2: Statistical modeling of giving choices

1 which is also in line with our findings in charity similarity

4. The more efficient the alternative charity is, the more likely the users are to switch (denoted by variable EfficiencyLevel). In particular, a boost from 80 to 90% efficiency translates to ≈0.2 points.
5. We did not observe a statistically significant effect for the locations of charities matching.

Overall, the results show that participants respond to the efficiency measure; when presented with an alternative that spends more of the donation directly on program expenses (as opposed to fundraising and administrative expenses), they showed intend to switch. The likelihood of switching increases when the difference in efficiency is larger and when the alternative is more similar to the focal charity, i.e. has high cosine similarity. The experiment does not reveal a significant location effect, which is in conflict with our previous finding and our intuition. One explanation is that workers were not explicitly told to consider location (as not to bias) and thus might have thought they were supposed to ignore it. Further, it may be difficult to display locational preferences for a place one is not from. We note, however, that while there was a strong home bias in online donor behavior, a majority of donations were out-of-state. Thus it is plausible that the participants of our survey do not consider location to be a strong factor. We believe that future experiments that involve real donations can help provide a more definitive explanation for this finding.

Based on these findings, we define the *Data-driven model* as follows: A charity c_j being a good substitute for another charity c_i is assigned a probability $p_{i,j}$:

$$p_{i,j} = \frac{\beta_0 + \beta_1 x_{i,j,1} + \beta_2 x_{i,j,2} + \beta_3 x_{i,j,3}}{5} \qquad (1)$$

where β_m values for $0 \leq m \leq 3$ are the weights presented in rows 1,2,3, and 6 in Table 2, $x_{i,j,1}$ is 1 if the location of c_i and c_j match and is 0 otherwise, $x_{i,j,2}$ is the difference in efficiency of charities c_i and c_j and finally $x_{i,j,3}$ is the cosine similarity between charities c_i and c_j.

Having defined choice models of varying constraints, we next estimate the inefficiencies in the current charity marketplace under these choice models.

Understanding marketplace efficiency via Choice Models

In this section we address the question of marketplace efficiency using the quantitative measures we constructed in the previous sections. Questions of this flavor are notoriously difficult to answer because they require an understanding of how people would change their decisions if they were made aware of available alternatives. By construction these "counterfactual choices" are not directly observed. In our setting, if we observe a particularly wasteful charity continue to attract user interest and donations, then it is tempting to conclude this is inefficient. While this certainly seems to qualify as inefficient in the conversational sense of the term, in the technical sense it is only inefficient if the donors of this charity would switch to other organizations if given the right information and opportunity.

The strategy we pursue is to propose a series of choice models in order to provide informative bounds on marketplace efficiency. We use the concept of "required similarity," a threshold that determines if a donor would switch to a more efficient charity if given the information and opportunity. With a loose threshold, many charities form acceptable substitutes and thus there is likely an efficient alternative for poorly run organizations. As required similarity becomes stricter, these sets shrink, often to a singleton. We present a wide range of tolerances, which usefully bound efficiency and also allows a reader to find the estimate that corresponds to her prior on donor choice behavior. We define *loss* as the efficiency difference between the focal charity and the most efficient charity in the qualifying set.

1. **Qualifying Set for Models 1-5:** The assumption on choice behavior we are making is that a donor would be willing to switch their contribution to any more efficient charity within the qualified set. If there are few acceptable substitutes, then relatively high over-head costs can be readily explained. This framework is meant to simulate a world in which people read a charity's mission statement and observe a high-level summary of a charity's books in a frictionless market[5]. To do so, we make a number of simplifying assumptions that are unlikely to hold in practice. Nonetheless, we note that the similarity thresholds are calibrated with data on the perceived similarity.
2. **Qualifying Set for Model 6:** Based on price, similarity and location effects estimated in the previous section, we estimate the marketplace efficiency compared to a world where every donation goes through a recommendation system that provides up to 3 alternatives before a donor makes a charitable contribution to a charity of their choice; where the 3 charities are chosen such that:
   (a) Alternative charities have higher efficiency compared to the focal charity.
   (b) Alternative charities are sorted by the predicted likelihood of switching based on equation 1.

[5]The donors know about all in the qualifying set or that there is a system that tells them the best in group and the user follows that deterministically.

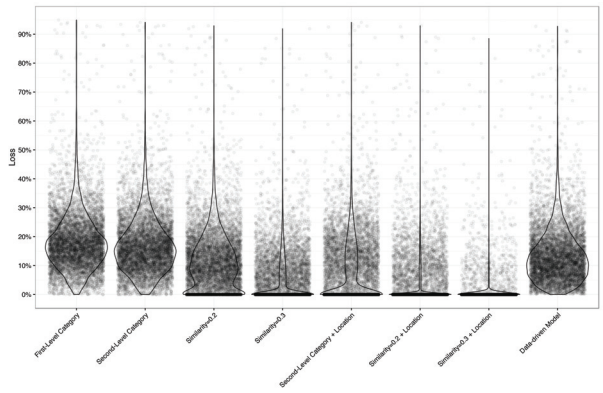


Figure 6: Loss distributions under various choice models.

(c) Alternative charities are presented one at a time.[6]

In order to estimate marketplace efficiency, we first calculate the likelihood that a donor who is about to donate to charity c_i would switch her donation to charity c_j if provided the necessary information for each c_i and c_j such that overhead of c_i is higher than overhead of c_j. Having computed $p_{i,j}$ for each charity pair, we simulate donor choices when a recommendation system presents them 3 alternative charities with the highest $p_{i,j}$ values and identify the alternative that would have been chosen along with the efficiency gain associated with the switch.

In Figure 6 we look at the distribution of loss in the current charity marketplace under 8 choice models: *first-level category, second-level category, similarity=0.2, similarity=0.3, second-level category + location, similarity=0.2 + location, similarity=0.3 + location,* and *data-driven model*. For each model, we present a violin graph to visualize the efficiency loss distribution. In addition to the violin plot, we include data points for each charity for deeper inspection. If the required similarity is given by the first- or second-level category then loss seems quite high, over 15% average loss and a substantial part of the distribution with about 30% loss. Requiring cosine distance of 0.2 shifts the loss distribution significantly toward zero, but still displaying substantial loss. The real change occurs when requiring either a cosine distance of 0.2 + location or distance of 0.3. In these cases the modal loss is zero, indicating there is not an "acceptable alternative" for most charities. Still, there is a long tail meaningful loss for all thresholds. Also note that the inefficiencies estimated based on the Data-driven model are in between the two extremes. There is a substantial mass around 10-30% loss compared to similarity=0.2 + location, but this mass is smaller compared to the first- and second-level category estimates.

Table 3 gives total loss in efficiency measured in dollars using the contributions data, which tell the same story. The unrealistically high loss estimated by using CN categories drops dramatically to 3.7 billion dollars with similarity 0.2 + location. With our strictest requirement, loss is less than a billion dollars. Loss estimated under the data-driven model is approximately 10.6 billion dollars, or just over 9% of total contributions for the charities we consider. This estimation is based on the assumption that the users trust the recommendation system and do not ignore its suggestions. Future work to build trustworthy and user-friendly recommendation systems is vital to achieve such efficiency gains.

[6]This is also equivalent to a setting where the alternatives are presented in one page but the donor makes decision independently and in the order presented.

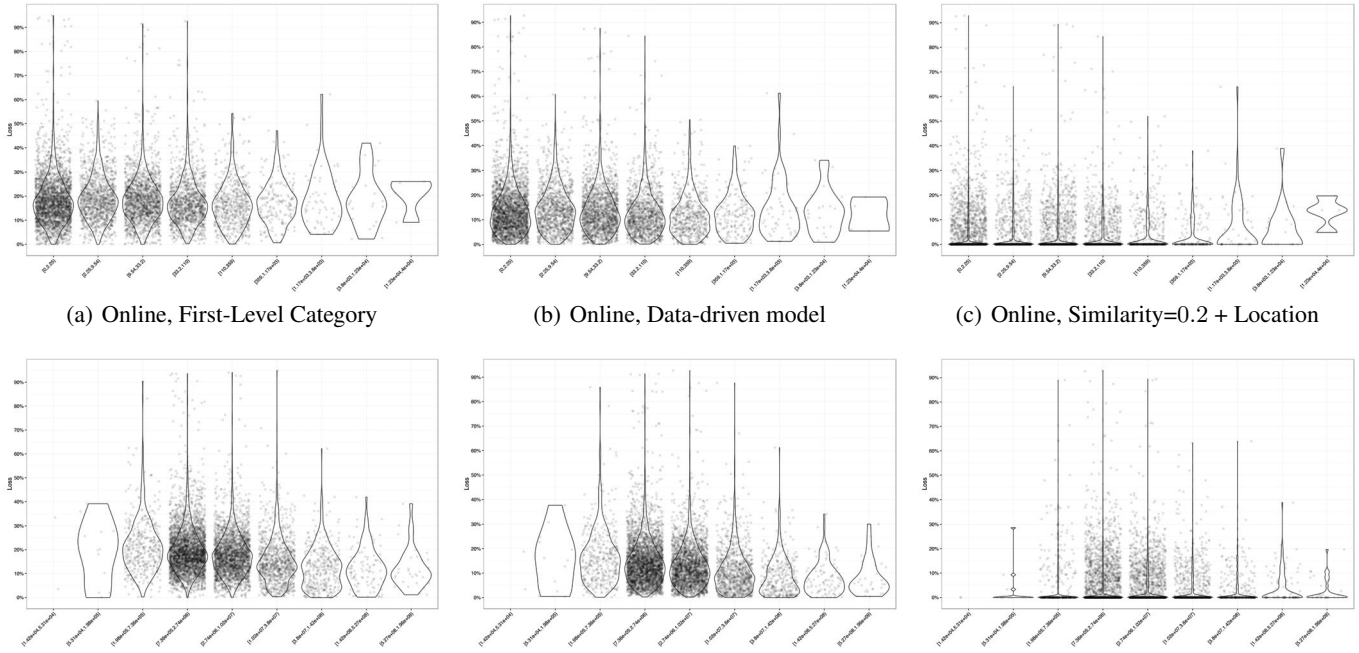
	Substitution Model	Loss Sum $ Billions
1	First-Level Category	15.4
2	Second-Level Category	13.6
3	Second-Level Category + Location	7.2
4	Similarity=0.2 + Location	3.7
5	Similarity=0.3 + Location	0.9
6	Similarity=0.2	9.8
7	Similarity=0.3	4.0
8	Data-driven Model	10.6

Table 3: Loss estimated based on different choice models

In Figure 7 we look at the loss distributions across charity sizes (log binned) for total online donors and overall contributions. An interesting pattern emerges for (7-a,b,c)—larger charities exhibit higher loss distributions[7]. This is rather surprising given that overhead and size are negatively correlated–smaller organizations tend to have larger overheads (Figure 2(b)). The intuition of this finding, upon reflection, is readily understood. Small charities are more likely to fill a particular niche and thus less likely to have acceptable substitutes. Large charities, in contrast, have either broader aims or address an issue that receives attention from a greater number related organizations. The takeaway is that if consumers were to make informed choices, the biggest dollar impact would come from the large, inefficient organizations.

A few points are worth noting. First, the seemingly large inefficiencies go away when stricter similarity, especially location, is required. Given the "home bias" observed earlier and similarity judgments of works, these requirements are very likely relevant for some donors. Second, at all levels of required similarity, there are inefficient laggards driving loss up. Third, at an intermediate threshold such as similarity 0.2 + location, the median and modal loss are both very low, but there is a meaningful segment of contributions that have loss above 10%. Larger losses are predicted under the data-driven model, which indicates that a wide adoption of a recommendation or comparison tool could meaningfully impact marketplace efficiency. Finally, since the loss was generally higher for larger charities, an important source of efficiency gains would come from donations redirected from large, high-overhead organizations.

[7]While this pattern is not observed in 7-d,e, it is observed in 7-f.



(a) Online, First-Level Category (b) Online, Data-driven model (c) Online, Similarity=0.2 + Location

(d) Overall, First-Level Category (e) Overall, Data-driven model (f) Overall, Similarity=0.2 + Location

Figure 7: Charity marketplace inefficiencies computed under various choice models for online donors and overall contributions

Discussion and Future Directions

The goal of this work was to improve our understanding of the marketplace for charitable giving. To do so, we used novel data sources and computational techniques. To the best of our knowledge this is the first work to take this approach. Our analyses are made possible by a few institutional features that make these charities amenable to study using computational techniques: 1) charities must file IRS 990 forms that detail their costs, revenues, and charitable mission 2) these forms are publicly available and aggregated by sites like Charity Navigator. We further complement these data with web browsing logs and crowdsourcing techniques.

We first gave a high level overview of the market through the lens of these data. We then defined a simple measure of similarity between charitable organizations and validated it through an Amazon Mechanical Turk survey. Importantly, this measure captures a much finer-grained characterization of an organizations' purpose as compared to simple categories given by Charity Navigator. The similarity scores allow us to simulate different choice models in order to study marketplace efficiency. Loss in the system due to poorly run charities varies widely depending on the required level of similarity. It is disturbingly high when using Charity Navigator's categories, but drops by 75% with an intermediate similarity threshold and a location requirement. Nonetheless, at all thresholds there is still a long tail of inefficient organizations that have close substitutes. Further, the offending organizations tend to be much larger than average. To provide a more concrete estimate we characterize price sensitivity in charitable giving through a Mechanical Turk experiment and demonstrate how a simple recommendation system can save billions of dollars for consumers.

A few caveats are worth keeping in mind. First, we cannot directly observe the set of charities that any individual donor would truly find to be acceptable substitutes. Charities may meaningfully differ in ways that are not captured by their mission statement or location and the importance of such characteristics presumably varies across donors. Our approach is to vary a set of assumptions to provide informative bounds. We complement these assumptions with a hypothetical survey and the estimates based on these data lie within the endpoints of the assumption-based bounds. Second, our data-driven estimates as based on a Mechanical Turk study. Mechanical Turk has a diverse labor force that is generally representative of the population of US Internet users (Ipeirotis 2010) and produce high quality work (Mason and Suri 2012). However, these workers are not necessarily representative of people that participate in the marketplace for charitable giving. Third, the existing evidence indicates that people give for a host of reasons beyond pure altruism (a genuine concern about the cause), such as social image concerns (Lacetera and Macis 2010), "warm glow" (Andreoni 1990) and to appease requests to give (Andreoni, Rao, and Trachtman Forthcoming; Castillo, Petrie, and Wardell 2014). Givers motivated purely by these factors may not care about charity quality, but these studies also reveal that many givers are motivated by a genuine desire to help others. Fourth, giving efficiency and effectiveness are different concepts—a charity could deliver on its mission to complete a given project, but there may be better projects that achieve their goals at lower cost or with higher probability (Karlan and Wood 2014). Surfacing ef-

fectiveness information, where available, could deepen the notion of efficiency we use in this paper. Finally, if many donors started making informed choices, charities would presumably react in a host of ways. We have not attempted to model these reactions—our work here is taken from the perspective of the consumer holding other factors fixed. Given these limitations, we view our estimates of inefficiency as informing the questions posed at the outset but not providing a definitive answers.

Future work will be instrumental in exploring questions raised by these caveats. A natural step would be to construct a functioning recommendation system to use in a randomized controlled trial with real money on the line. An application of such a recommendation system would be in giving "portals," such as those used in employer matching gifts programs. A well-functioning system would require a deeper understanding about the features that have the largest impact on choices and how feature weights vary across people. Further, computational methods can improve and expand our ability to measure accountability, transparency measures (e.g. are the tax returns prepared by an independent accountant?) and effectiveness (are the projects described known to be effective) to provide a more holistic view of this marketplace. These future directions highlight the potential impact projects in this space have to not only extend our understanding of the charity marketplace but also help people make more informed choices and thus improve welfare.

References

Andreoni, J., and Payne, A. 2003. Do government grants to private charities crowd out giving or fund-raising? *American Economic Review* 93(3):792–812.

Andreoni, J.; Rao, J. M.; and Trachtman, H. Forthcoming. Avoiding the ask: A field experiment on altruism, empathy, and charitable giving.

Andreoni, J. 1990. Impure altruism and donations to public goods: A theory of warm-glow giving. *The economic journal* 100(401):464–477.

Andreoni, J. 2006. Philanthropy. In Kolm, S., and Ythier, J., eds., *Handbook on the Economics of Giving, Reciprocity and Altruism*. Elsevier. 1201–1269.

Auten, G. E.; Clotfelter, C. T.; and Schmalbeck, R. L. 2000. Taxes and philanthropy among the wealthy. *Does atlas shrug* 392–424.

Auten, G. E.; Sieg, H.; and Clotfelter, C. T. 2002. Charitable giving, income, and taxes: an analysis of panel data. *American Economic Review* 371–382.

Baeza-Yates, R.; Ribeiro-Neto, B.; et al. 1999. *Modern information retrieval*, volume 463. ACM press New York.

Bakija, J. M.; Gale, W. G.; and Slemrod, J. B. 2003. Charitable bequests and taxes on inheritances and estates: Aggregate evidence from across states and time. *American Economic Review* 93(2):366–370.

Bennett, R., and Savani, S. 2003. Predicting the accuracy of public perceptions of charity performance. *Journal of Targeting, Measurement and Analysis for Marketing* 11(4):326–342.

Buchheit, S., and Parsons, L. M. 2006. An experimental investigation of accounting information?s influence on the individual giving process. *Journal of Accounting and Public Policy* 25(6):666–686.

Castillo, M.; Petrie, R.; and Wardell, C. 2014. Fundraising through online social networks: A field experiment on peer-to-peer solicitation. *Journal of Public Economics* 114:29–35.

Clotfelter, C. T. 2001. Who are the alumni donors? giving by two generations of alumni from selective colleges. *Nonprofit Management and Leadership* 12(2):119–138.

Eckel, C. C., and Grossman, P. J. 2008. Subsidizing charitable contributions: a natural field experiment comparing matching and rebate subsidies. *Experimental Economics* 11(3):234–252.

Glaser, J. S. 1994. *The United Way scandal: An insider's account of what went wrong and why*, volume 22. John Wiley & Sons Inc.

Gordon, T. P.; Knock, C. L.; and Neely, D. G. 2009. The role of rating agencies in the market for charitable contributions: An empirical test. *Journal of Accounting and Public Policy* 28(6):469–484.

Hager, M., and Greenlee, J. 2004. How important is a nonprofit's bottom line? The uses and abuses of financial data. *Search of the Nonprofit Sector. Eds. Frumkin, P., Imber, JB, New Brunswick, NJ, Transaction* 85–96.

Ipeirotis, P. G. 2010. Demographics of mechanical turk. Technical report, Tech. Rep. No. CeDER-10-01. New York: New York University. Available: http://hdl.handle.net/2451/29585.

Karlan, D., and List, J. A. 2007. Does price matter in charitable giving? evidence from a large-scale natural field experiment. *The American Economic Review* 1774–1793.

Karlan, D., and Wood, D. H. 2014. The effect of effectiveness: Donor response to aid effectiveness in a direct mail fundraising experiment. Technical report, National Bureau of Economic Research.

Lacetera, N., and Macis, M. 2010. Social image concerns and prosocial behavior: Field evidence from a nonlinear incentive scheme. *Journal of Economic Behavior & Organization* 76(2):225–237.

Mason, W., and Suri, S. 2012. Conducting behavioral research on amazon?s mechanical turk. *Behavior research methods* 44(1):1–23.

Parsons, L. M. 2007. The impact of financial information and voluntary disclosures on contributions to not-for-profit organizations. *Behavioral research in accounting* 19(1):179–196.

Randolph, W. C. 1995. Dynamic income, progressive taxes, and the timing of charitable contributions. *Journal of Political Economy* 709–738.

Weisbrod, B. A., and Dominguez, N. D. 1986. Demand for collective goods in private nonprofit markets: Can fundraising expenditures help overcome free-rider behavior? *Journal of public economics* 30(1):83–96.