Contents lists available at ScienceDirect

Research Policy

journal homepage: www.elsevier.com/locate/respol

Research note

Crowdfunding and regional entrepreneurial investment: an application of the CrowdBerkeley database[☆]

Sandy Yu[a], Scott Johnson[b], Chiayu Lai[c], Antonio Cricelli[d], Lee Fleming[b,d,*]

[a] Carlson School of Business, University of Minnesota, Minneapolis, CA, United States
[b] Fung Institute for Engineering Leadership, University of California, Berkeley, CA, United States
[c] Department of Information Management, National Sun Yat-Sen University, Taiwan, ROC
[d] School of Business, University of California, Berkeley, CA, United States

ARTICLE INFO

Keywords:
Crowdfunding
Entrepreneurship
Regional economics

ABSTRACT

Crowdfunding platforms enable individuals to solicit small investments, donations, or loans over the Internet from a wide variety of funders; they have emerged as a new and potentially important source of funds for entrepreneurial and philanthropic initiatives. We build and present three databases for public use, including Kickstarter, Kiva, and CrowdRise, and link regional measures of Kickstarter to entrepreneurial ventures listed in Crunchbase. We find that Kickstarter projects in a region correlate with increased angel investing activity, even after instrumenting with projects that should not be of interest to investors. The paper describes scraping tools, database schema, descriptive statistics, dashboards, access for research and policy use, and general reflections on building open databases for the research community.

1. Introduction

It has been traditionally challenging for startups to attract external financing. The usual sources are risk-averse bankers and conventional business loans or equity capital, the latter provided by small groups of sophisticated investors who invest in return for a share of the venture. Thus, many new ventures remain unfunded. Recently, entrepreneurs have used the Internet platforms to appeal to the “crowd”; by listing and describing their investment or cause, entrepreneurs can reach a large audience where each individual provides a small amount. Crowdfunding (hereafter CF) platforms bypass standard financial intermediaries and enable founders to directly solicit money for a variety of for-profit, artistic, or social projects, often but not always in return for future products or possibly equity. These projects vary greatly in size and goal. They can be local art projects requiring a few hundred dollars, social projects to fundraise for a cause asking for a few thousand dollars, or entrepreneurs seeking hundreds of thousands of dollars to fund their startup using CF as an alternative to traditional venture capital financing (Mollick, 2014).

In recent years, crowdfunding as a method of entrepreneurial financing has grown very quickly. In 2009, there were 53 platforms worldwide that raised approximately $1 billion. In 2014, there were over 750 platforms that raised approximately $12 billion. 2015 saw an estimated raise of $33B and in 2016, CF was expected to surpass Venture Capital investment (Massolutions, 2015). This growth of the CF market occurred despite an uncertain political landscape in the US where the JOBS Act (which legalized equity CF) passed in April 2012 and the SEC only fully legalized equity CF on May 16, 2016.

Crowdfunding platforms have become diverse and specialized and target increasingly differentiated segments. Typologies have proliferated; here we organize into four categories, including debt-based, charity, rewards-based, and equity. Debt-based CF (the most popular by dollar volume, see Gray and Zhang, 2017) has attracted increasing attention from traditional finance and is part of the emerging FinTech sector. It is often called peer-to-peer/P2P lending or marketplace lending; prospective borrowers list their requirements and investors can choose whether to accept the credit terms. Loans utilizing debt-based CF are often for personal reasons such as debt consolidation or home improvement. Prominent examples include Prosper and LendingClub. Charity CF is very similar to traditional charitable fundraising, where individuals donate to a project or cause for individuals or organizations. Examples include unexpected medical bills or fundraising for team-based marathons. Two of the largest charity CF platforms are GoFundMe and DonorsChoose. Rewards-based CF allows individuals to

[☆] The Kauffman Foundation made the CrowdBerkeley database possible. The work was also supported by NSF grants #1536022 and 1661311, the Haas School of Business, and the Coleman Fung Institute for Engineering Leadership; errors and omissions remain the authors.

[*] Corresponding author.
E-mail addresses: sandyyu@berkeley.edu (S. Yu), scott.johnson@berkeley.edu (S. Johnson), chiayu06@gmail.com (C. Lai), tony@haas.berkeley.edu (A. Cricelli), lfleming@berkeley.edu (L. Fleming).

http://dx.doi.org/10.1016/j.respol.2017.07.008
Received 8 March 2017; Received in revised form 17 July 2017; Accepted 19 July 2017
Available online 07 September 2017

Table 1
Selected summary statistics scraped from Kickstarter website April 2009 to end of December 2016. Data are slightly greater than Mollick (2014) and appear to include more failed campaigns (confirmed in personal communication with Ethan Mollick on August 9, 2016).

	N	Mean	Min	Max	Std Dev
Goal (USD)	312,594	43323.85	0.01	169732132.02	1100695.00
Amount pledged (USD)	312,594	8941.48	0.00	20338986.27	91806.67
Backer count	312,594	107.07	0.00	219382.00	956.51
Comment count	312,594	38.22	0.00	389373.00	1164.05
Campaign duration	312,594	34.50	1.00	92.00	13.04
Has video	312,594	0.71	0.00	1.00	0.45

fund a project in return for a reward, which can range from a token of appreciation, such as credits in a movie, to a product or service, such as a beta-version of a product. This form of CF allows entrepreneurs to raise money without incurring debt or sacrificing equity. This is the most widely-known type of CF and examples include Kickstarter and Indiegogo. Equity CF is most akin to angel and VC financing, where individuals contribute money in return for shares of a company. These companies are still early in their lifecycle. Examples of equity crowdfunding platforms include AngelList and CircleUp. Equity CF appears to have accounted for 7.35% of the total global crowdfunding industry in 2015 (Massolutions 2015), hence, most investors do not receive equity.

The Fung Institute at UC Berkeley, with the support of the Kauffman Foundation, has assembled a publicly-available database on three CF platforms to date: Kickstarter, Kiva, and CrowdRise. Kickstarter is the largest rewards-based crowdfunding website by traffic, number of backers, and total dollars pledged (Massolution) and a global industry leader. Kickstarter claims (midway through 2017) to have raised over $3 B since its founding in April 2009, and these successes have made the website a popular platform for the study and analysis of rewards-based CF. Kiva operates as a non-profit with a mission to fund loans that alleviate poverty. CrowdRise provides a platform for philanthropic fund raising without expectation of payback to funders.

The motivation behind developing these databases is to provide researchers and policy makers with a comprehensive summary of projects that is as accurate and current as possible (the websites are ideally scraped and updated daily). Each project page contains a description of the project, funds raised, rewards offered, project backers, comments, and updates. The panel data contains daily statistics of number of backers, amount funded, number of comments, and number of updates for each project while it is live. Kickstarter does not provide full statistics on backers, due to privacy concerns. Due to financial constraints, we unfortunately provide data only on U.S. platforms, though CF is quite popular around the world (Gray and Zhang, 2017). We provide a MetaBase interface for users who are unfamiliar with how to access SQL databases and summary statistics and graphical illustrations.

To provide an example of the types of research that the databases enable, the paper explores the impact of Kickstarter campaigns upon regional entrepreneurial funding, by linking Crunchbase data to Kickstarter. It appears that Kickstarter campaigns, and technology campaigns in particular, correlate with an increase the number of angel funding rounds in a region. The note will conclude by reflecting on the challenges of building a database for the research community. Appendices include technical details of scraping, database schema and updating, and user access.

2. The Databases: Kickstarter

Kickstarter (KS) is one of the largest rewards-based CF platforms and includes projects from a diverse set of categories, including technology, food, design, and games. KS data is scraped from publicly

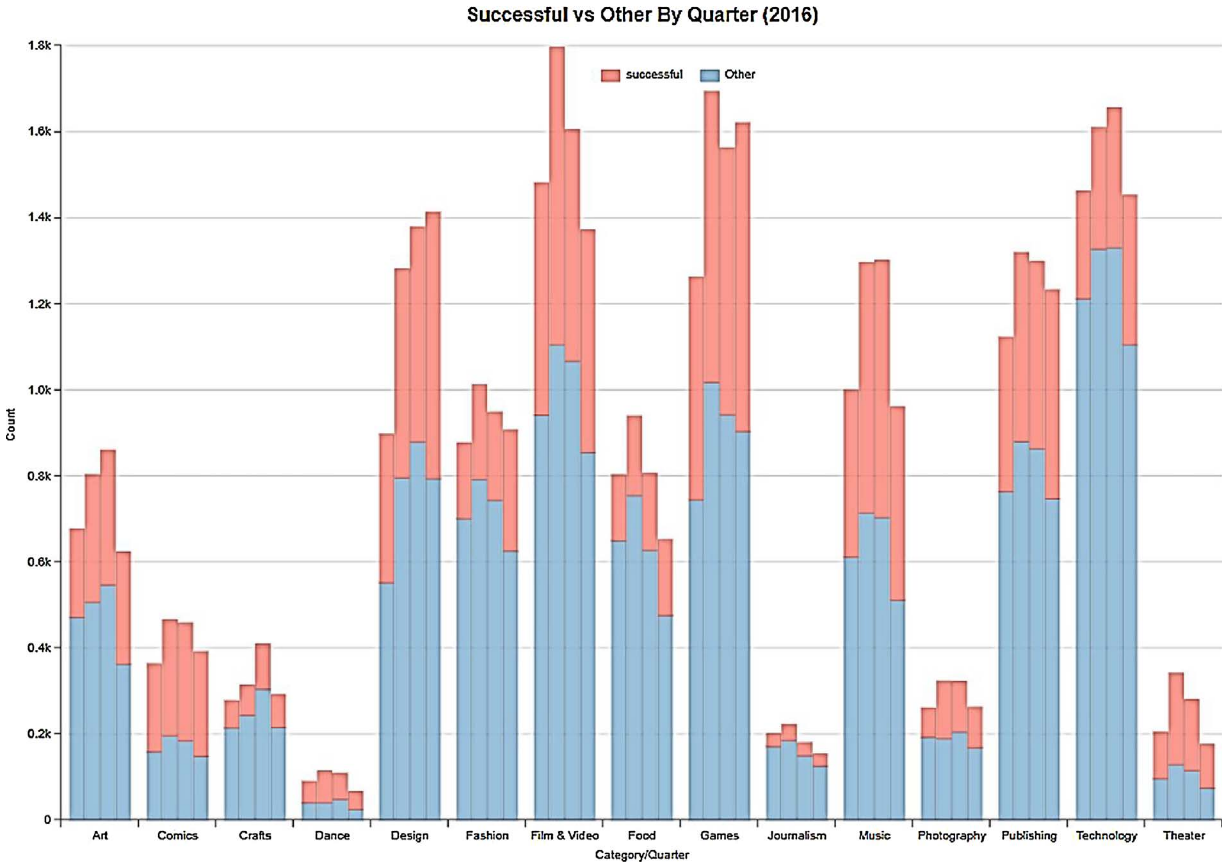


Fig. 1. Number of successful and failed Kickstarter campaigns by quarter for all categories in 2016. The campaigns appear seasonal and that seasonality varies slightly by category.

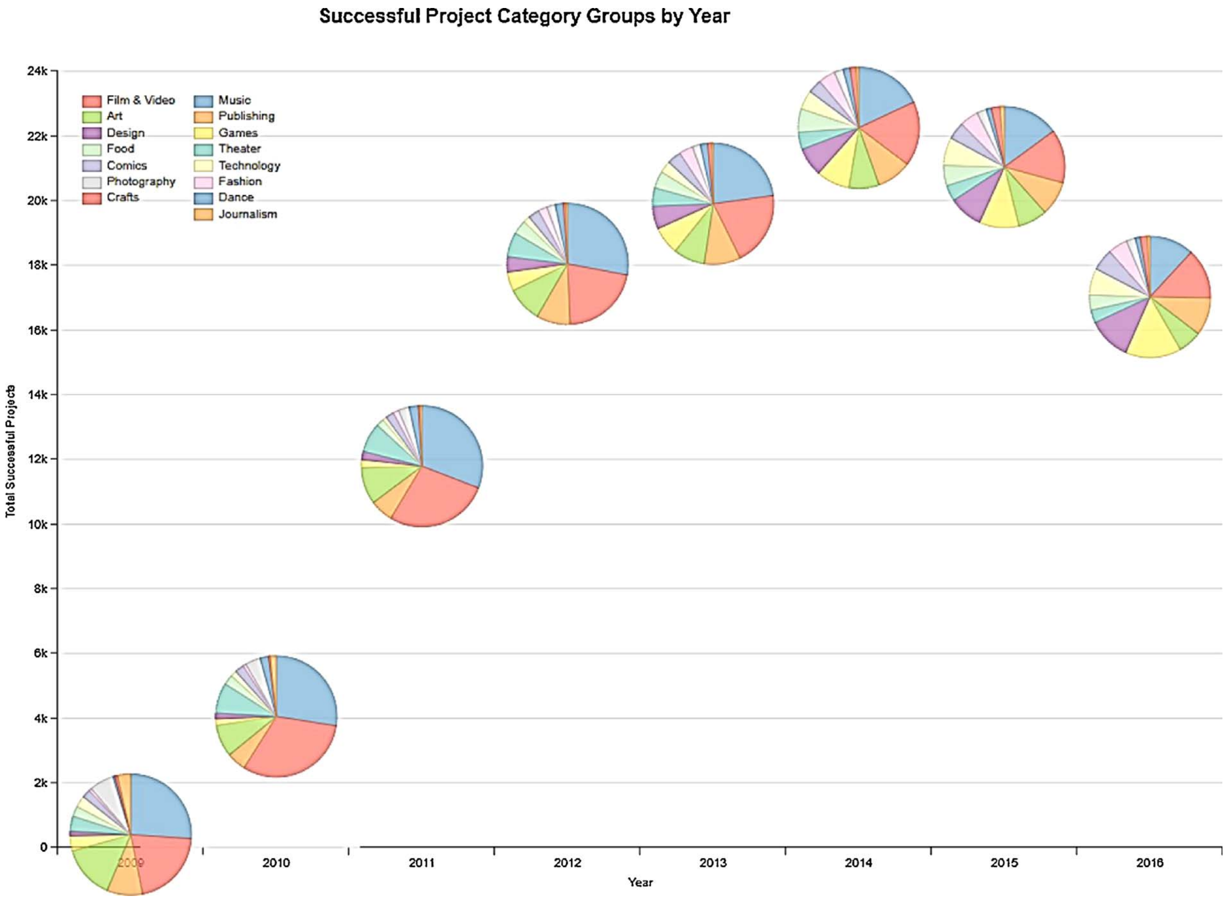


Fig. 2. Proportions of successful Kickstarter projects since 2009. Crafts and music have become relatively smaller proportions of projects while technology has become a larger proportion.

Table 2
Selected summary statistics scraped from Kiva website, April 2005 to end of December 2016.

	N	Mean	Min	Max	Std Dev
Funded amount (USD)	1,310,658	780.91	0	100000.00	969.00
Lender count	1,310,658	22.28	0	2986.00	26.76
Loan amount (USD)	1,310,658	823.76	0	100000.00	1007.20
Repaid amount (USD)	33,212	721.16	0	50000.00	883.02

accessible project pages starting from 2009 and contains ~312,000 campaigns (as of the end of December 2016). The collected variables include project title, description, location (city, state, country), founder details, fundraising goal amount (in USD), actual fundraised amount (in USD), category, and project status (success, failed, canceled). Further details on backers, such as comments, are also collected. Selected summary statistics are listed in Table 1 and Figs. 1 and 2.

3. The Databases: Kiva

Kiva is an international nonprofit that provides a CF platform to fund loans to borrowers in developing economies. Kiva lenders crowdfund on average $2.5 million in loans each week; more than one million loans have been funded. Loan amounts start at $25 increments and lenders are repaid over time. The Kiva data is scraped from publicly accessible loan pages starting from 2005 and contains ~1,310,000 observations (as of the end of December 2016). Collected variables on the loan-level include borrower and loan description, borrower location, number of lenders, total loan amount (in USD), and loan status. On the lender-level, the lender ID, location, loan date, and number of total loans is collected. Table 2 and Fig. 3 illustrate.

4. The Databases: CrowdRise

CrowdRise provides a CF platform for charitable and personal causes. Example projects include fundraising for charities, medical expenses, personal emergencies, and volunteer projects. CrowdRise data is scraped from publicly accessible campaign pages starting from 2010 and includes ~491,000 projects (as of April 2017). The collected variables include project name, description, organizer, fundraising goal (in USD), fundraised amount (in USD), donation dates, and donor comments. Table 3 and Fig. 4 illustrate.

5. Does Crowdfunding increase regional entrepreneurial funding?

Economic inequality has become a defining controversy of our time (Piketty and Goldhammer, 2014) and seemingly fueled populist reactions around the world, including within the U.S. The differential impact of technological change is cited as one potential cause of this inequality. Recent technological and social change, in the form of the Internet and rise of online CF communities, could increase or decrease this inequality. It might decrease regional inequality, if it increases innovation and entrepreneurship in regions away from traditional hubs such as Boston and Silicon Valley. It could also increase regional inequality if it drains resources from poorer regions as crowds become more aware of distant opportunities and send their money to wealthier regions. Here we provide an example of how these databases might be applied by investigating if Kickstarter activity in a region leads to an increase or decrease of entrepreneurial investment, as observed by

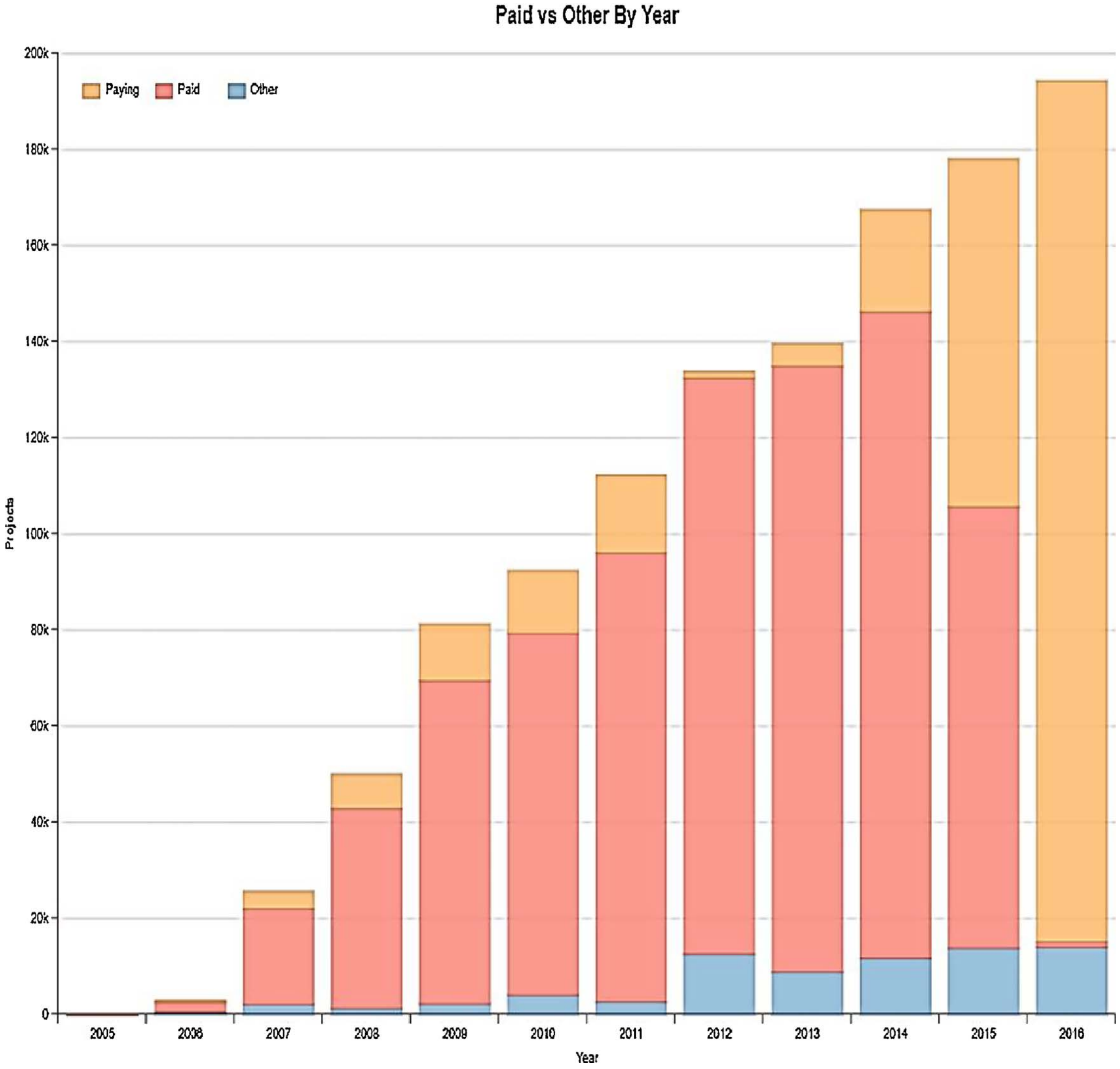


Fig. 3. Status of repayment of Kiva loans by year.

Table 3
Selected statistics scraped from CrowdRise website, Jan 2010 to end of April 2017.

	N	Mean	Min	Max	Std Dev
Amount raised (USD)	491,721	2185.81	0.00	100005654.00	153731.54
Donation count	491,721	10.06	0.00	23932.00	95.85
Team members	491,721	0.40	0.00	9011.00	14.15

Angel funding rounds in Crunchbase. This becomes an interesting question, as CF appears to be relatively stronger in regions with less venture capital funding, compared with traditional hubs such as Silicon Valley and Boston (Sorenson et al., 2016).

Crunchbase is an open source database maintained by TechCrunch, a leading technology news site. Although it is open source, Crunchbase has partnerships with ~900 venture capital firms and AngelList to ensure their public data is accurately represented. Crunchbase tends to have more early stage companies, which makes it ideal for examining nascent ventures and new venture formation. Crunchbase data includes founder profiles, company location (city, state, country), founding date, business description, funding milestones (date and amount), investors, and operational status (active, acquired, closed, IPO). For this example, company-level data is aggregated to the county level to examine new venture activity in regions across the U.S. In particular, the number of rounds of angel funding in a region is regressed upon the number of successful Kickstarter campaigns in that region.

Crowdfunding might decrease subsequent angel funding in a region if entrepreneurs substituted crowdsourced investment for angel investment. Alternately, CF might increase subsequent funding if early investors looked for ideas – and the validation of ideas –from CF success. Through CF platforms, investors can 1) gain more information about market traction and are 2) able to access more deals. Furthermore, several well-known angel investors have become active in investing in crowdfunded products or services after successful campaigns (Schroter, 2014). Teasing out these consistent mechanisms empirically strikes us as a fruitful direction for future research.

The relationship between CF and investment is difficult to establish with correlations, as many factors might influence both the number of CF campaigns and angel investments in a region. Fixed effects models could account for some of these factors, such as relatively static variables such as education levels of a workforce, geographical or institutional influences, or even wealth, assuming these do not change quickly over time. Other co-varying factors could change simultaneously, however; for example, the economic cycle could encourage both CF and angel investments.

We employ an instrumental variables approach to ameliorate these concerns (as illustrated below, non-instrumented models show similar though often attenuated relationships). Based on lexical similarity in Kickstarter projects and venture capital investments from VentureXpert, we divide Kickstarter projects into three categories: 1)

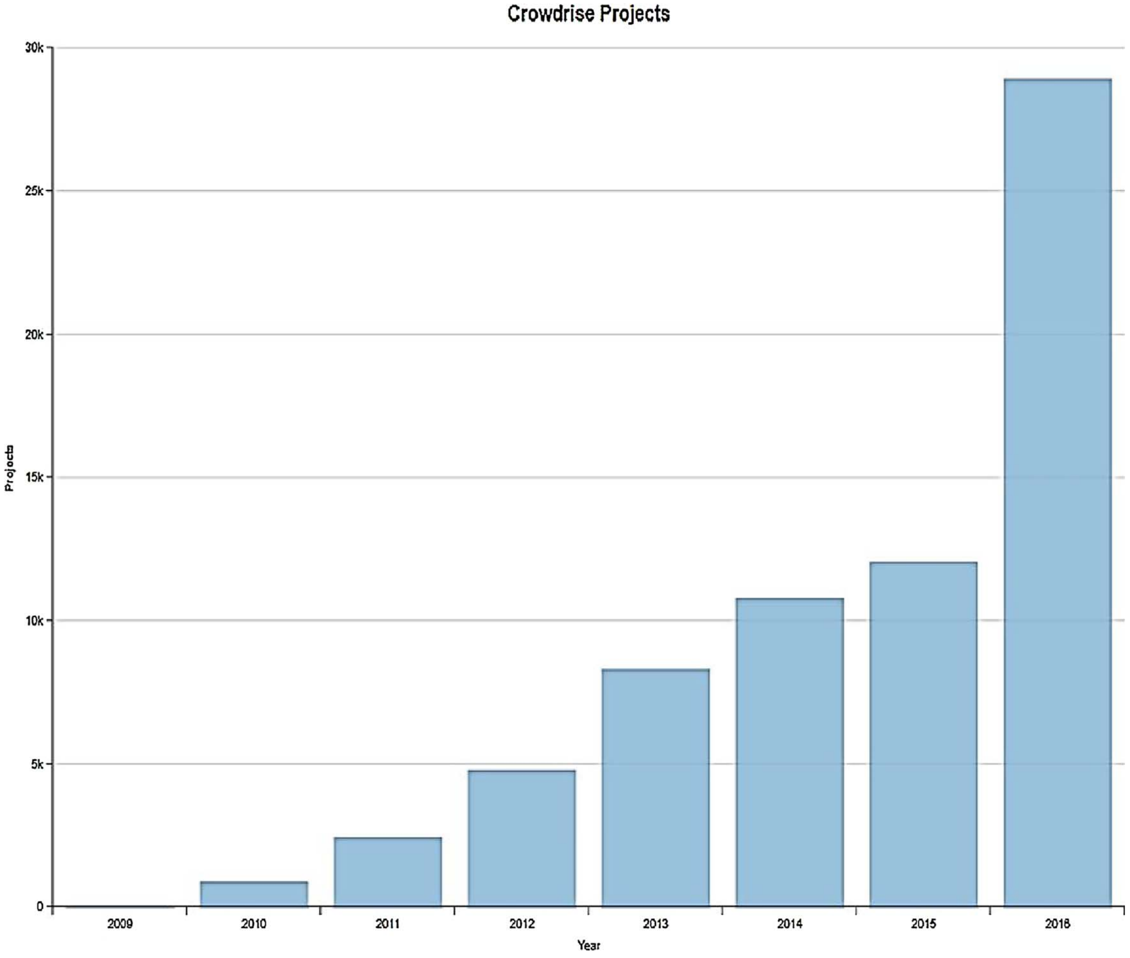


Fig. 4. Number of CrowdRise projects by year.

those of great interest to investors, 2) those of moderate interest, and 3) those of little interest. Fig. 5 illustrates (from Sorenson et al., 2016). Dark rows indicate Kickstarter campaigns with strong lexical overlap (i.e., similar words in their descriptions, see actual overlap in percentages) and include games, food, technology, fashion, crafts, and journalism. Gray rows indicate campaigns of little lexical overlap and include film and video, music, comics, and dance. Note the lack of dark entries in the VentureExpert Biotechnology and Medical/Health columns, which corroborates Kickstarter's prohibition of biotechnology and medical campaigns. Note also that this assumes that VentureExpert and Crunchbase investors have similar investment interests. Fig. 6 shows the contribution of technology campaigns, campaigns with strong lexical overlap (games, food, fashion, crafts, and journalism), and campaigns with little lexical overlap over time (film and video, music, comics, and dance).

We use the categories with little overlap as an instrument for categories of greatest overlap and do not consider those in the intermediate category. The logic of the exclusion restriction is that the projects that are not of interest to investors will correlate with the projects that are of interest, and yet attract no investment and therefore have no impact on subsequent entrepreneurial investment in the region. (See Supplementary Materials to Sorenson et al., 2016 for details.) The Cragg-Donald F statistic is 5999.56, indicating a very strong instrument. The amount of patenting and citations to patents in a region (in a particular year) control for the number and quality of available ideas.

Table 4 provides descriptive statistics and 5 considers the relationship between successful Kickstarter projects and the number of angel investments in a region, including year and region fixed effects (models that consider the amount of angel funding return similar results). An increase in successful Kickstarter projects correlates with angel investments and imposing a time trend indicates that the effect has been increasing over time (though the underlying effect loses significance with inclusion of the interaction effect in Table 5). Table 6 illustrates increased and significant effects when considering only successful Kickstarter technology campaigns. Fig. 7 interprets and graphs effect sizes. It would appear that Kickstarter crowdfunding draws greater entrepreneurial investment to a region, does not act as a substitute for angel investment, and that these trends have increased over time. By the end of the time period, it appears that a 1% increase in successful technology campaigns corresponds to an increase of over 0.4% in angel investments.

6. Reflections and opinions on building databases for the research community

Given the big investment required to build a database, and the many and often unanticipated questions that might be answered with it, it is a waste of research investment not to share it widely. Here we reflect on this process and offer suggestions for those who are considering contributing such a database. Making data public also reflects a now widespread trend across all sciences in making data fully accessible, and in particular, data needed to replicate published findings (King, 1995).

Finding the financial support to build databases is non-trivial and finding support to host them even more difficult. While most funding entities support and even require data sharing, reviewers often seem

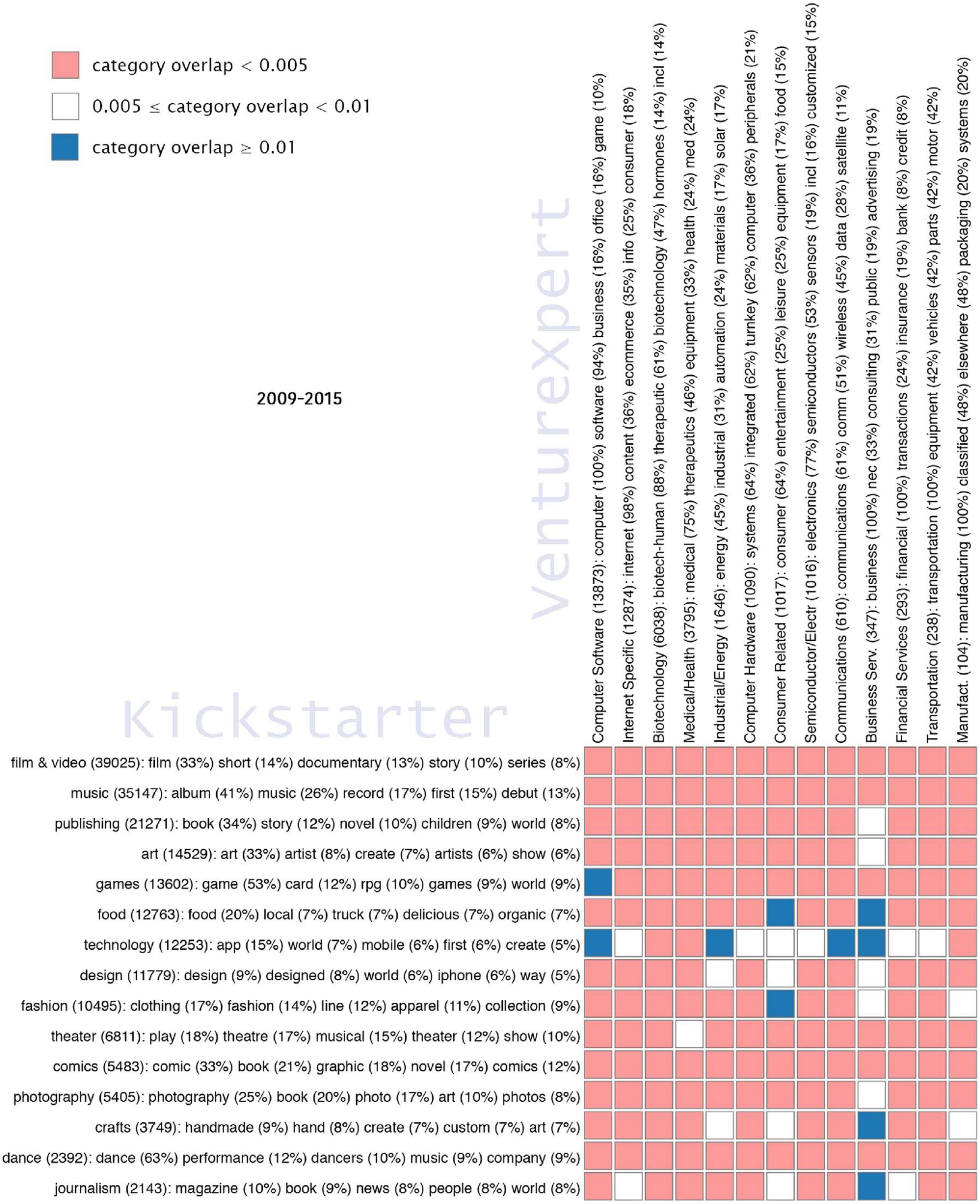


Fig. 5. (from Supplementary Materials, Sorenson et al., 2016): Lexical overlap between categories of Kickstarter and VentureExpert. Dark rows indicate strong overlap and are used as an independent variable; gray rows only indicate weak overlap and are used as an instrument. Remaining rows are not used.

reluctant to fund research proposals that ask for resources to document and make data widely available. Proposals that seek to maintain a previously developed database are particularly unpopular, even though interfaces break, data become outdated, and formats change (scraping websites for data is particularly vulnerable to changes in HTML code). The authors can offer little advice in this regard, except to ask that reviewers view such requests in a more positive light, and agencies perhaps allocate some fixed percentage of their support to such efforts.

Finding database builders is also no easy task. It is rare that a social scientist has the ability *and* interest in building a complete database and document how to use it. Students (often computer science or engineering majors) are often hired to program and build databases, however, they are usually temporary, rarely understand the context, and approach the problem without an understanding of social science methods. They require a great deal of attention and direction, particularly when it comes to testing and documentation (in particular, it is very difficult for such students to spot very obvious errors – simple things that would jump out to a social scientist).

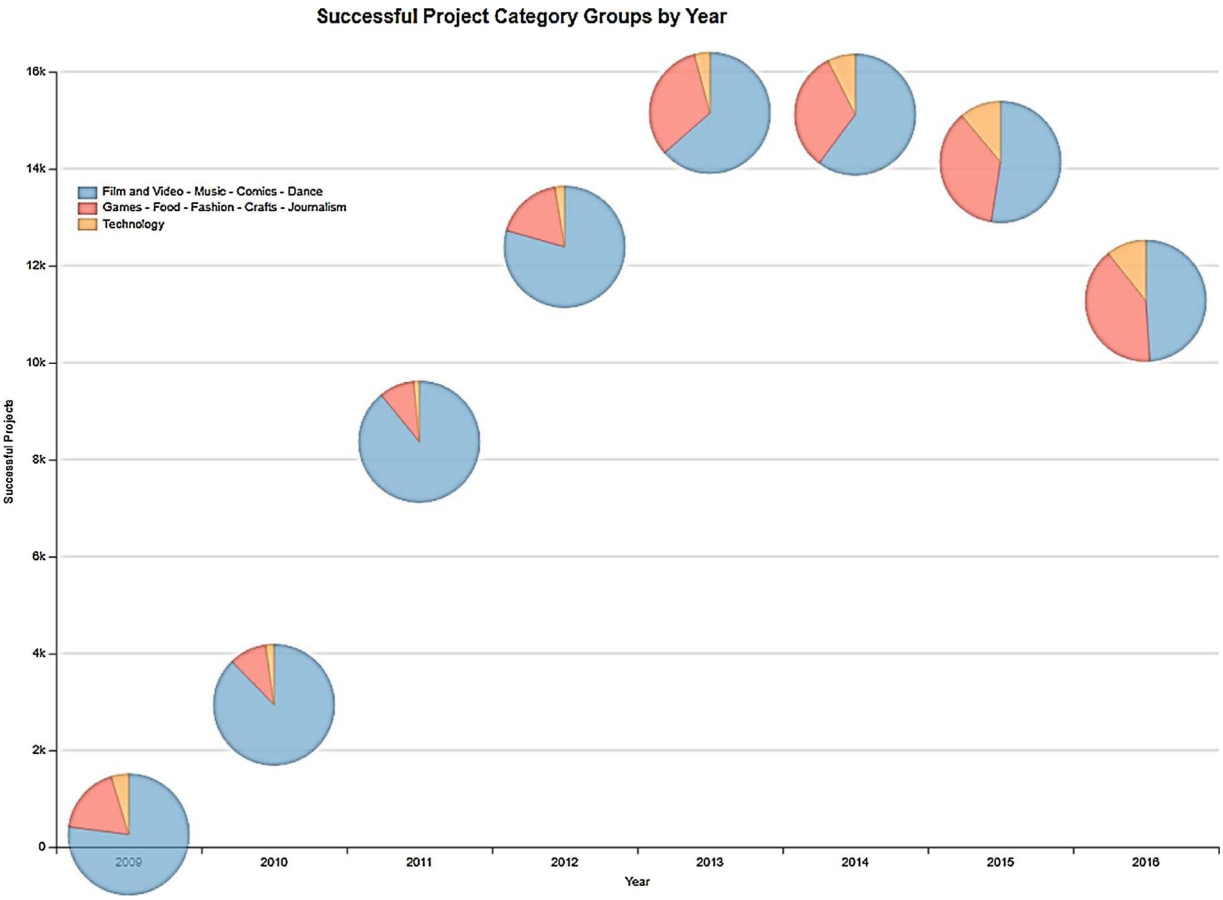


Fig. 6. Proportions of successful Kickstarter projects by category groups since 2009. Games, food, fashion, crafts, and journalism are categories with strong lexical overlap; film and video, music, comics, and dance are categories with weak lexical overlap. The overall contribution of game and technology campaigns are increasing over time.

Table 4
Descriptive statistics for main variable.

	Obs	Mean	Std. Dev.	Min	Max
Number of successful campaigns	22,575	0.70	6.26	0.00	281.00
Number of successful campaigns (technology only)	22,575	0.12	1.48	0.00	74.00
Number of angel investments	22,575	0.10	1.55	0.00	71.00
Number of patents	22,575	42.43	369.31	0.00	25956.00
Number of citations	22,575	757.36	6803.46	0.00	426996.00

Errors are endemic to databases and often seem to multiply like weeds once you start looking. Users have little tolerance for them and will sometimes malign the probity of the author, even when the data were shared with the best of intentions. Errors arise from no end of sources: bad original data (which then get blamed on the author), changes in data format (often undocumented by the original source), changes in website architecture which then breaks a scraping tool, simple programming errors, size problems, and inadequate software and/or hardware. Users can be exceptionally helpful in debugging a database, though this requires an IT infrastructure and process for gathering, tracking, and acting on feedback (which is quite costly). This process of improving accuracy also tends to be very research question specific – a database that has proven accurate in one context rarely proves so in another. It is unfortunate but important to keep in mind that building a general database is a much deeper and more onerous task than building a database for a narrow set of questions. Users should view shared data as the starting point which enables them to build their own database, not as a ready to analyze off the shelf product.

Once the database is built, the author needs to find a server to host it. We would advise finding Information Technology (IT) professionals to do this. University IT departments often have the capacity to do this and can provide firewalls if the author wishes to ask for a login to track usage (this Crowdfunding database is hosted, for example, by the Haas School of Business) or protect sensitive or proprietary data (in particular, care must be taken not to reveal human subjects or financial data). Harvard's Institute for Quantitative Social Science (IQSS) hosts Dataverse (http://dataverse.org/), a very successful and popular site for sharing databases. The Dataverse also allows easy posting of documentation and accompanying papers, so that authors are more likely to earn deserved citation credit for their contribution.

One last bit of warning to those who offer up a database for general consumption. You will receive plaintive emails, some offering very reasonable and useful feedback, and some asking for help across a variety of topics such as opening a file or writing a dissertation. These requests are best received with appreciation and patience, respectively.

7. Conclusion – and possibilities

We have scraped, built databases, and provided public interfaces for three prominent Crowdfunding platforms, including Kickstarter, Kiva,

Table 5
Naïve regressions (models 1-2), estimate of instrument strength (campaigns not of interest to Angel investors, model 3), and two-stage least squares instrumental variable regression estimates of the effect of successful Kickstarter campaigns on the number of future angel investments (models 4-5), with two-way (year and county) fixed effects for 3225 counties over 7 years.

	1) Count of Angel Investments	2) Count of Angel Investments	3) First Stage	4) Count of Angel Investments	5) Count of Angel Investments
KS Successful	0.0699***	−0.000962		0.111***	0.0308
	(0.00235)	(0.00612)		(0.00487)	(0.0215)
KS Successful x time		0.0147***			0.0258***
		(0.00117)			(0.00518)
IV = KS non-commercial			0.455***		
			(0.00588)		
Patents	0.00545***	0.00922***	−0.0496***	0.00749***	0.0111***
	(0.00168)	(0.00170)	(0.00447)	(0.00171)	(0.00214)
Citations	0.000786	0.00273**	-0.0371***	0.00257**	0.00474***
	(0.00115)	(0.00116)	(0.00307)	(0.00118)	(0.000932)
Constant	0.00926***	0.000413			
	(0.00246)	(0.00255)			
Observations	22,575	22,575	22,575	22,575	22,575
Cragg-Donald Wald F-Stat			5999.56		
R-squared	0.049	0.057		0.034	0.043
Number of FIPS	3,225	3,225	3,225	3,225	3,225

Table 6
Naïve regressions (models 1-2), estimate of instrument strength (campaigns not of interest to Angel investors, model 3), and two-stage least squares instrumental variable regression estimates of the effect of successful technology Kickstarter campaigns on the number of future angel investments (models 4-5), with two-way (year and county) fixed effects for 3225 counties over 7 years.

	1) Count of Angel Investments	2) Count of Angel Investments	3) First Stage	4) Count of Angel Investments	5) Count of Angel Investments
KS Successful (Tech only)	0.137***	0.130***		0.327***	0.229***
	(0.00429)	(0.0130)		(0.0148)	(0.0542)
KS Successful (Tech only) x time		0.00140			0.0372***
		(0.00242)			(0.0120)
IV = KS non-commercial			0.155***		
			(0.00350)		
Patents	0.00622***	0.00636***	-0.0310***	0.0121***	0.0143***
	(0.00168)	(0.00169)	(0.00267)	(0.00181)	(0.00239)
Citations	7.71e-05	0.000124	−0.0148***	0.00328***	0.00462***
	(0.00115)	(0.00115)	(0.00183)	(0.00123)	(0.000986)
Constant	0.0103***	0.0100***			
	(0.00244)	(0.00249)			
Observations	22,575	22,575	22,575	22,575	22,575
Cragg-Donald Wald F stat			1961.7		
R-squared	0.056	0.056		−0.040	0.035
Number of FIPS	3,225	3,225	3,225	3,225	3,225

and CrowdRise. The underlying databases are in SQL and will be updated daily; more databases may follow. To illustrate how such databases might be used, we established that the number of angel funding rounds in a region correlates with the number of Kickstarter campaigns. To strengthen causal inference, we applied a lexical overlap method that separated campaigns of interest to investors (such as technology) from those of little or no interest (such as arts and philanthropy). From these instrumented regressions, it appears more likely that technology campaigns have a strong, positive, and recently increasing impact on angel funding in a region.

The databases provide a number of future research opportunities. For example, CF campaigns might improve the quality of entrepreneurial ventures and/or select out the best opportunities. If the wisdom of the crowd evaluates projects correctly (Mollick and Nanda, 2015), the quality of new firms, as measured by subsequent financing or the proportion of successful firms, should increase within a region. Founding rates may drop, but ultimate success rates may go up. Alternately, CF activity might decrease the average quality of new ventures because the financial barriers to entry become too low. If the crowd is not adept at evaluating projects with financial potential, very low quality projects may get funded by a CF platform. If this is the case, one might expect lower amounts of subsequent financing and higher failure rates within a region. CF activity probably impacts particular industries more heavily. Furthermore, if CF is indeed successful in encouraging entrepreneurship in a region, then one would expect the distribution of firms to change in that region over time.

Finally, does CF funding go to richer or poorer regions, and if so, where does that funding come from (Burtch et al., 2014)? Are richer regions sources — or sinks — of CF funding? If CF flowed from rich to poor regions, it could be a viable mechanism to decrease regional inequality. On the other hand, if we observed money flowing from poor to rich regions, this would appear to heighten inequality. This mechanism may vary by the definition of rich and poor (for example, population, per capital income, distance from an urban center). One can estimate dyadic models of flows between all pairs of counties, controlling for distance and other observed covariates.

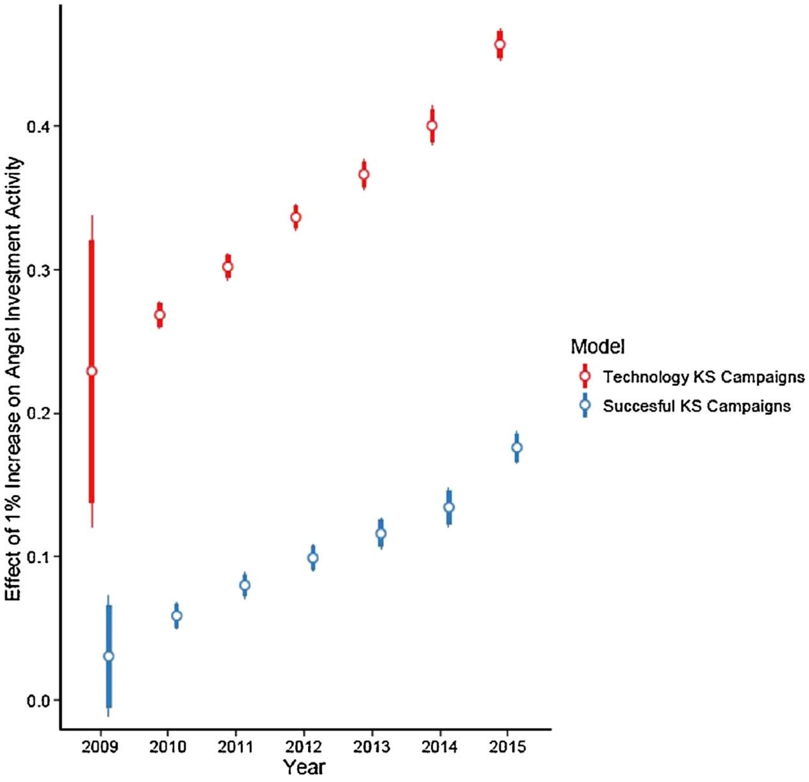


Fig. 7. Instrumented impact of 1) all successful Kickstarter campaigns on subsequent entrepreneurial firm starts in a region and 2) just successful technology campaigns.

Appendices

Accessing CrowdBerkeley Databases

The main website for CrowdBerkeley is http://www.crowd.berkeley.edu and the databases can be accessed at: https://crowdfunding.haas.berkeley.edu/wp/. To download selections from the database, simply register an email address. Once logged in users can submit SQL queries to the database under the "Scraped Public Database" tab (see Fig. 8). This tab also has a drop-down menu with sample SQL queries. To allow access without a knowledge of SQL, we use Metabase, located at http://fung-datascience.coe.berkeley.edu/. Accessing the data requires an email address ending in @gmail.com.

Coding details

Initial database setup

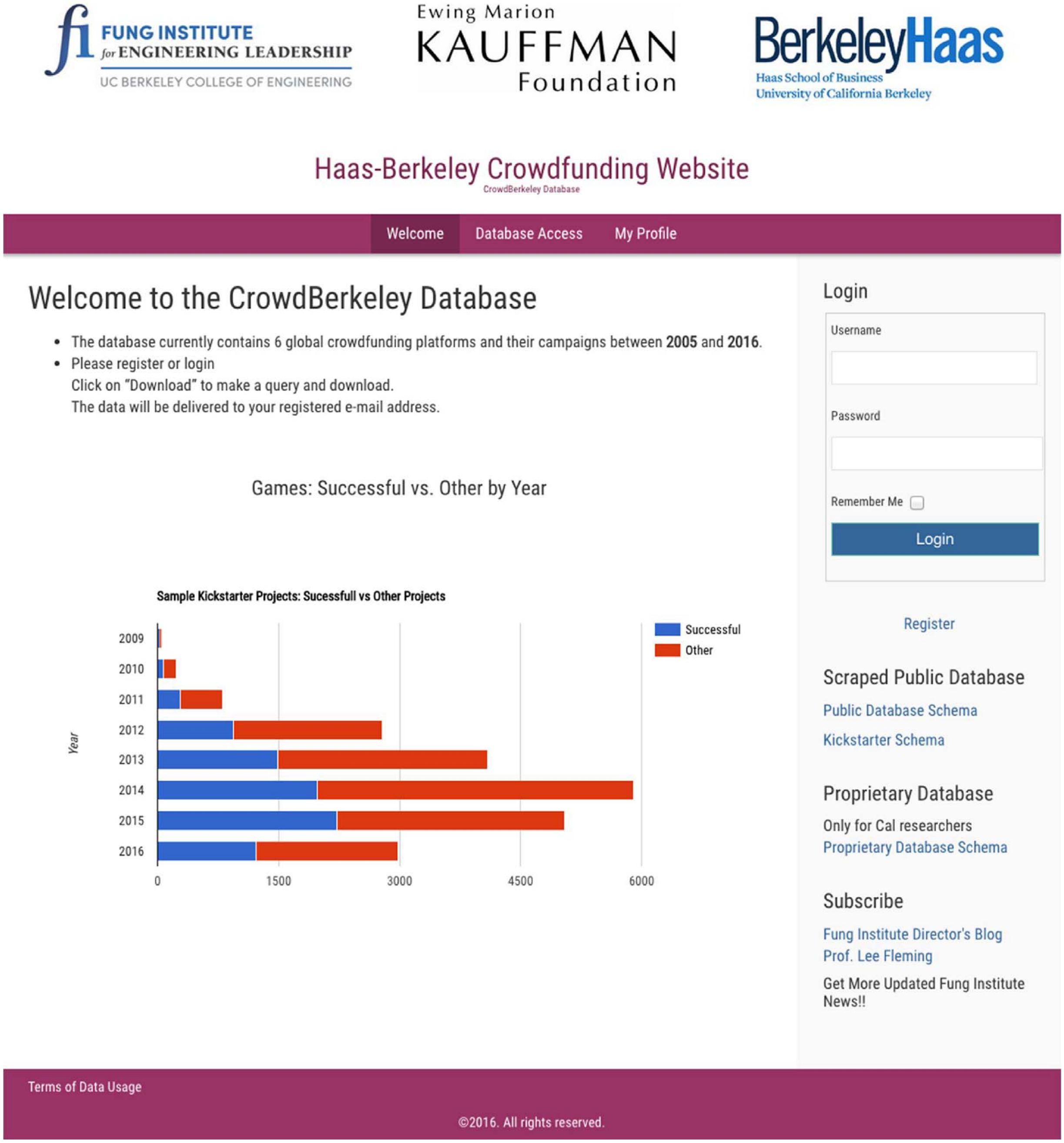
The basic pipeline for populating the database was to first download the complete HTML for projects into an intermediate database, write scrapers to parse the raw HTML, and extract the desired information to the main database. This process is illustrated in Fig. 9 and is described in detail below. All code for downloading and parsing the HTML was written in Python.

The URLs for each project are needed in order to scrape the project data. Our initial source of URLs was Webrobots.io, a website which provides various scraping and crawling services. Their data contained the URLs from almost every project on Kickstarter as of their last scrape, as well as location information for each project. The location data was stored in the location table in the main Kickstarter database, whereas the URLs were sent to the intermediate database. The HTML Downloader took these URLs and downloaded the root page, as well as the page for updates, description, and rewards. For example, if the URL https://www.kickstarter.com/projects/ysnet/shenmue-3 is in the intermediate database, the HTML from the following pages would be added to the intermediate database:

- https://www.kickstarter.com/projects/ysnet/shenmue-3
- https://www.kickstarter.com/projects/ysnet/shenmue-3/updates
- https://www.kickstarter.com/projects/ysnet/shenmue-3/description
- https://www.kickstarter.com/projects/ysnet/shenmue-3/rewards

The Project HTML Parser then took the HTML from the intermediate database and put as much relevant information as possible into the main database (see below for descriptions of variables).

The last step in the initial setup was getting all the comments; these could not be downloaded in the same way because not all comments load at once. Using the HTML Downloader, the HTML from the comments page of each project in the Kickstarter database that had one or more comment was sent to the Comment HTML Parser. The comment parser sent additional comment pages to the HTML Downloader, as needed. Once all the comments from a project were downloaded, they were parsed and sent to the main Kickstarter database.



Haas-Berkeley Crowdfunding Website

CrowdBerkeley Database

Welcome to the CrowdBerkeley Database

- The database currently contains 6 global crowdfunding platforms and their campaigns between **2005** and **2016**.
- Please register or login
  Click on "Download" to make a query and download.
  The data will be delivered to your registered e-mail address.

Scraped Public Database

Proprietary Database

Only for Cal researchers

Subscribe

Get More Updated Fung Institute News!!

©2016. All rights reserved.

FUNG INSTITUTE for ENGINEERING LEADERSHIP
UC BERKELEY COLLEGE OF ENGINEERING
Ewing Marion KAUFFMAN Foundation
BerkeleyHaas
Haas School of Business
University of California Berkeley
Welcome
Database Access
My Profile
Login
Username
Password
Remember Me □
Login
Register
Public Database Schema
Kickstarter Schema
Proprietary Database Schema
Fung Institute Director's Blog
Prof. Lee Fleming
Terms of Data Usage
Games: Successful vs. Other by Year
Sample Kickstarter Projects: Sucessfull vs Other Projects
Year
2009
2010
2011
2012
2013
2014
2015
2016
0
1500
3000
4500
6000
Successful
Other

Fig. 8. Public data tab of database website. Users can enter customized SQL commands or choose from a drop-down menu of sample queries. Users must first register at initial website.

Database updates

During the daily update, only the URLs from live projects (projects that have not yet reached their funding deadline) need to be sent to the intermediate database. Most of these live project URLs come from our Kickstarter database. Projects that have been added to Kickstarter since the last update are found by scraping their new projects page (https://www.kickstarter.com/discover/advanced?sort=newest). This page requires a separate program to download and scrape it because it requires scrolling to load all the projects; the HTML Downloader can only access links. The new project page also contains the location data from these new projects, and this is added to the main database as well. The live project URLs are sent to the intermediate database, downloaded, and then scraped with the Project HTML Parser, like before. The comments from the live URLs are also downloaded, but the Comment HTML Parser will stop downloading the comments from a project once it gets to a comment already in the Kickstarter database. Since Kickstarter's new project page provides both URLs for projects not in the current database and location data for those projects, we no longer rely on Webrobots.io.

Updating the database daily enables us to produce panel data for projects, which are stored in the funding trend table of the main Kickstarter database. The funding trend table tracks a project's progress over time by capturing the daily changes to the amount of money raised, the number of backers funding the project, the number of comments and updates, and the fundraising status of the project. New rows are added to the funding trend table for ongoing projects and projects that have just ended at the end of each daily update.

A complete update adds the URLs from all projects from the Kickstarter database to the intermediate database, downloads, scrapes, and sends the data back to the database. While most data remains static after the end date of the project (e.g. fundraising data), some is not (e.g. new comments

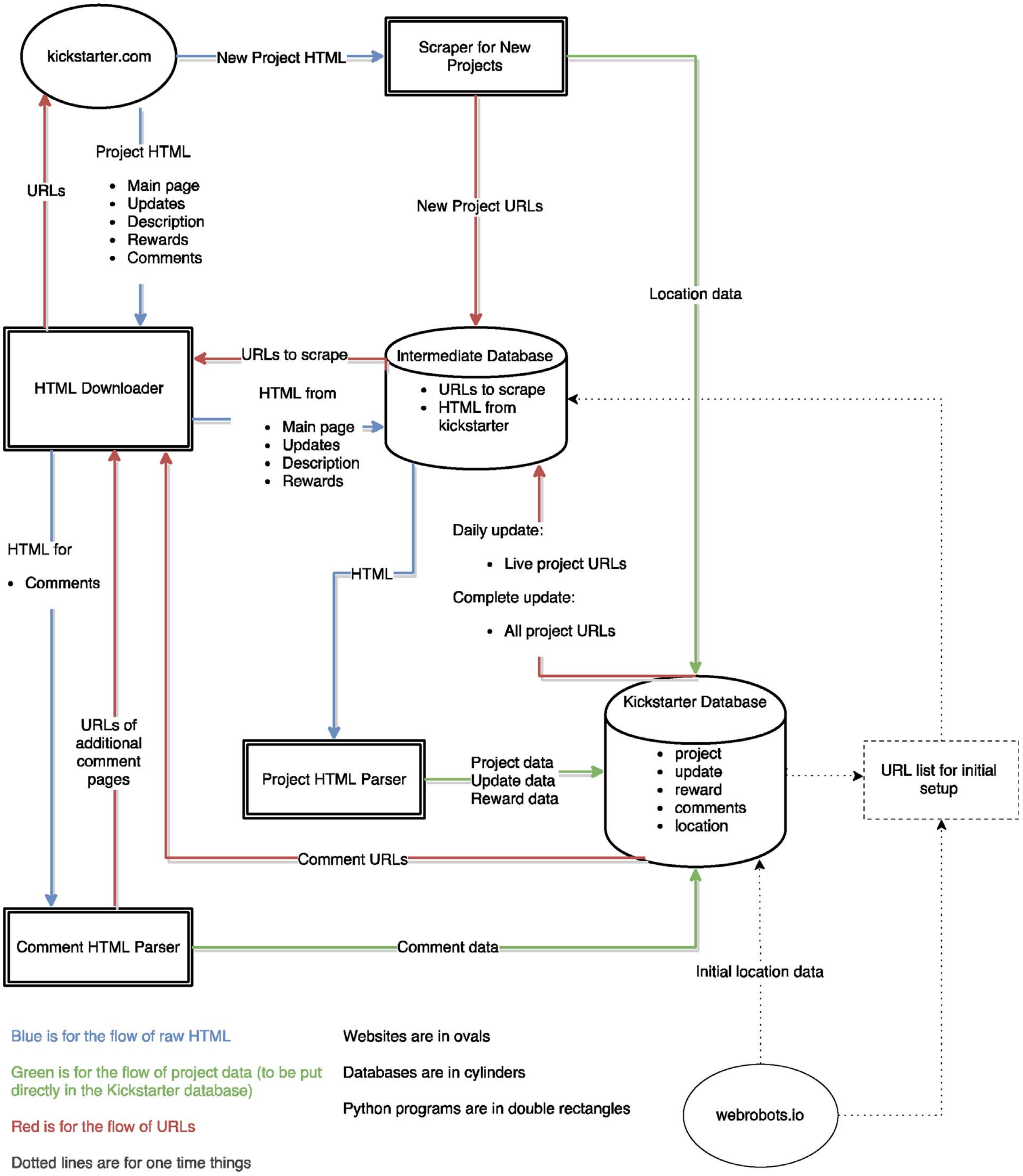


Fig. 9. Update process for the Kickstarter database.

and updates), and a complete update captures these changes. The most recent projects from Kickstarter are not scraped and the funding trend table is not updated, as this is best left to a daily update.

Database schema

The **Kickstarter** database contains six tables. The main table is the project table, which contains a row with a unique project *id* (primary key) for each project. *title* is the string title of the project. *description* is a short blurb written by the project creator about the project. Due to space considerations we do not store the full text project descriptions in the database, though we do have a number of other variables describing the contents of the description.

- *url* is the unique URL for the main page of the project
- *goal* is the fundraising goal amount
- *status* is a string describing the funding status of the project, e.g. “live”, “successful”, “cancelled”, etc.
- *amount_pledged* is the total amount of funds pledged towards the goal. If the project is live, this number will change, with daily changes

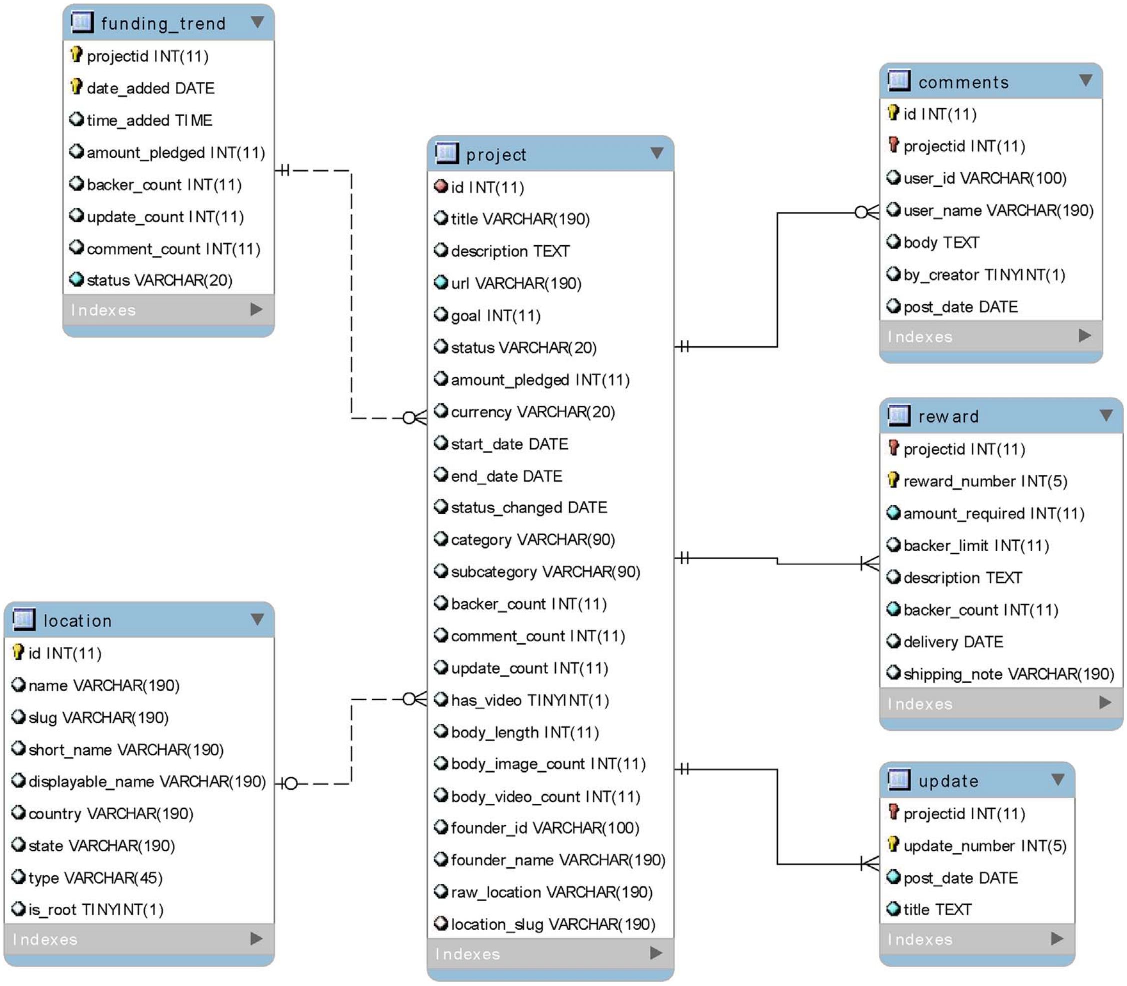


Fig. 10. Kickstarter database schema.

documented in the funding_trend table; otherwise, this number reflects the final amount raised for the project.

- *currency* describes the units of currency for *goal* and *amount_pledged*.
- *start_date* and *end_date* delimit the funding period for the project.
- *status_changed* is the date when the variable *status* changes, e.g. from “live” to “successful”.
- *category* is the broad category description of the project (e.g. “Food”, “Music”), out of 15 total available on Kickstarter, whereas *subcategory* is a specific category description (e.g. “Bacon”, “R & B”), out of 150 + available on Kickstarter.
- *backer_count* is the total number of users funding the project; this variable is again subject to change, with daily changes documented in *funding_trend*. Similarly, *comment_count* and *update_count* are the total number of comments and updates, respectively.
- *has_video* is a binary variable equal to 1 if the project has a central, explanatory video and 0 otherwise.
- *body_length* is the number of characters in the full text description.
- *body_image_count* and *body_video_count* are the number of images and videos embedded in the full text description, respectively.
- *founder_id* is the user id of the project creator; *founder_name* is the name of the project creator.
- *raw_location* is the project's location as it appears on the website; *location_slug* corresponds to an entry in the location table.

The comments, rewards, and update tables are connected to the project table by the project id and contains rows for each comment, update, and rewards, respectively. In the comment table, id} is the primary key for a comment. *projectid* is the *id* of the project to which the comment is posted (foreign key to the project table). *user_id* is the id of the poster of the comment; *user_name* is the name of the poster. *body* is the text of the comment. *by_creator* is a binary variable equal to 1 if the comment is written by the project creator and 0 otherwise. *post_date* is the date the comment was posted.

In the rewards table, *projectid* is again a foreign key to the project table. *rewards_number* is the order number of the rewards; for example, this variable would be equal to 2 if it corresponds to the second out of five possible rewards. *amount_required* is the minimum amount a user must pledge to receive the rewards. *backer_limit* is the maximum number of backers who may receive the rewards, in the case of limited supplies. *description* is the text description of the rewards. backer_count is the total number of backers who have pledged to receive the rewards. *delivery* is the date of delivery for the rewards, and *shipping_note* contains any addition information about shipping.

An update is a message posted by the project creator. While similar to comments, updates are stored on their own page on the website. They tend to be longer messages and are often only viewable by project backers. In the update table, *projectid* again corresponds to the project to which the update is posted. *update_number*, like *rewards_number* in the rewards table, is the order number of the update. *post_date* is the date the update was posted. *title* is the title of the update.

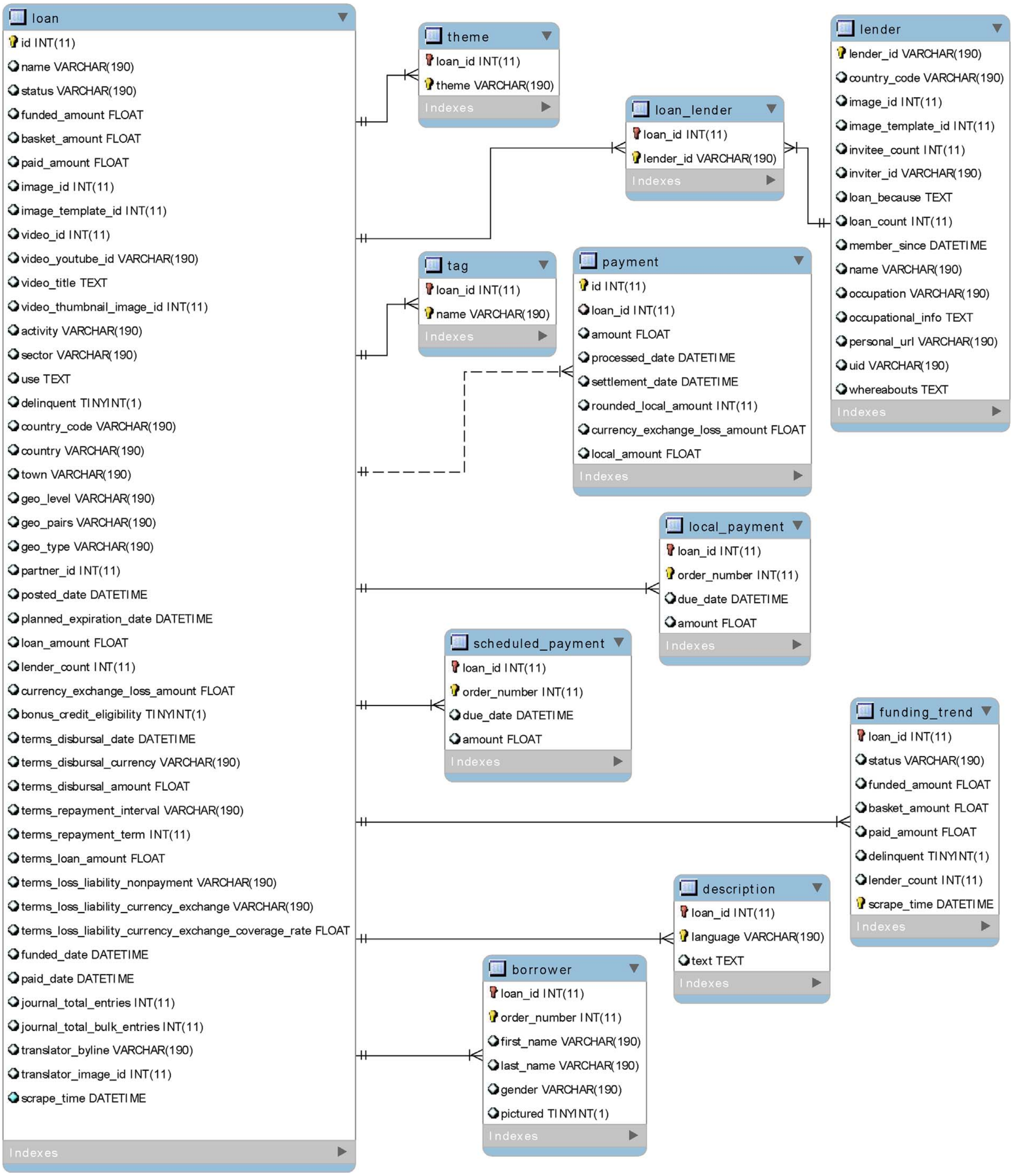


Fig. 11. Kiva database schema.

The funding_trend table contains panel data from projects over time and is also connected by project id. Each row in the funding_trend table reports the funding-related data at a specific time with new rows added daily. *projectid* is a foreign key to the project table. *date_added* and *time_added* describe when the row was added to the database, i.e. the time of the update. The variables *amount_pledged*, *backer_count*, *update_count*, *comment_count*, and *status* all correspond to the variables of the same name in the project table at the time of the update.

Finally, the location table contains location data, such as name, state, and country. *location_slug* corresponds to *location_slug* in the project database. See Fig. 10 for a pictorial representation of the database.

The **Kiva** database contains 11 tables (see Fig. 11). The *loan* table contains most of the information related to every loan made on Kiva, such as where the borrower is located, the how much is being asked for, and how much has been lended. Additional loan information can be found in the *description*, *payment*, *local_payment*, *scheduled_payment*, *tag, theme, loan_lender,* and *borrower* tables. The last two tables are *funding_trend* and *lender.*

- *loan* contains most of the information related to every loan made on Kiva, such as where the borrower is located, the how much is being asked for, and how much has been lent

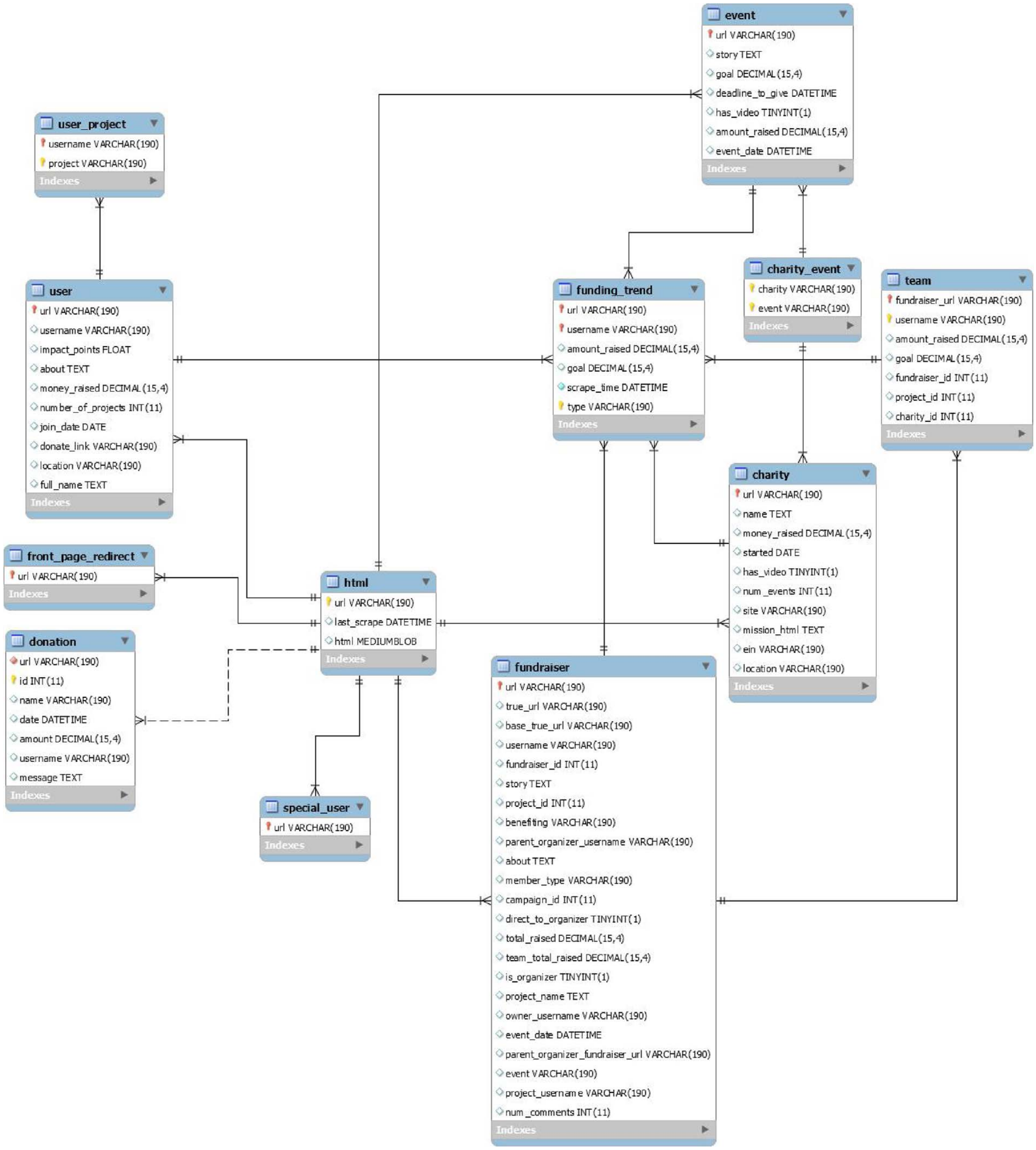


Fig. 12. CrowdRise database schema.

- *description* contains the description of each loan in at least one language
- *local_payment* contains the payment schedule from borrower to Field Partner in the local currency
- *scheduled_payment* contains the payment schedule from Field Partner to lenders in USD
- *tag* contains the tag(s) associated with each loan, such as “Schooling” or “Technology”
- *theme* contains the themes associated with each loan, such as “Health” or “Higher Education”
- *loan_lender* contains the lenders that have lent money to each loan contains information about the people that requested each loan
- *funding_trend* contains daily snapshots of the current loans, giving details such as *funded_amount* and *lender_count* over time
- *lender* contains information about every lender

The **CrowdRise** database contains 12 tables (see Fig. 12).

- *charity* contains the charities on CrowdRise, which raise money through events and fundraisers.
- *charity_event* contains the event(s) associated with each charity, if any.
- *donation* contains the donations made to each fundraiser with very approximate dates.

- *event* contains the events on CrowdRise, which raise money through fundraisers.
- *front_page_redirects* contains the urls of pages that redirected to the front page.
- *fundraiser* contains the fundraisers on CrowdRise. Fundaisers can have an associated event in the *event* column and/or an associated charity in the *benefiting* column.
- *funding_trend* tracks the amount raised and goal of charities, events, fundraisers, team members, and users.
- *html* contains the latest HTML downloaded for each url.
- *special_user* contains the users who got their own special page, e.g. www.CrowdRise.com/maccaxchallenge2013.
- *team* contains the users that are part of a fundraiser.
- *user* contains the users who have started a fundraiser.
- *user_project* contains each user's specific fundraiser page. For example, if www.CrowdRise.com/jonasbrothers is main fundraiser, then the fundraiser page for anniemarshal is www.CrowdRise.com/jonasbrothers/fundraiser/anniemarshal.

References

Burtch, G., Ghose, A., Wattal, S., 2014. Cultural Differences and geography as determinants of online prosocial lending. MIS Q. 38 (3), 773–794.

Gray, M., Zhang, B., 2017. Crowdfunding: understanding diversity. https://www.repository.cam.ac.uk/handle/1810/263730.

King, G., 1995. Replication, Replication. PS: Politi. Sci. Polit. 28 (September), 444–452.

Massolutions, 2015. Crowdfunding Industry Report. . http://crowdexpert.com/crowdfunding-industry-statistics/.

Mollick, E., Nanda, R., 2015. Wisdom or Madness? Comparing Crowds with Expert Evaluation in Funding the Arts. Manage. Sci. 62 (6), 1533–1553.

Mollick, E.R., 2014. The Dynamics of Crowdfunding: An Exploratory Study. J. Bus. Venturing 29 (January (1)), 1–16.

Piketty, T., Goldhammer, A., 2014. Capital in the Twenty-first Century. Belknap Press, Cambridge.

Schroter, W., 2014. Meet 7 Angel Investors Who Love Crowdfunding. Forbes. https://www.forbes.com/sites/wilschroter/2014/05/02/meet-7-angel-investors-who-love-crowdfunding#15807f383096.

Sorenson, O., Assenova, V., Li, G., Boada, J., Fleming, L., 2016. Expanding innovation finance via crowdfunding. Science 354 (Dec (6319)), 1526–1528.