Woods, Clinton; Yu, Han; Huang, Hong

Article

Predicting the success of entrepreneurial campaigns in crowdfunding: a spatiotemporal approach

Journal of Innovation and Entrepreneurship

Provided in Cooperation with:
Springer Nature

Suggested Citation: Woods, Clinton; Yu, Han; Huang, Hong (2020) : Predicting the success of entrepreneurial campaigns in crowdfunding: a spatiotemporal approach, Journal of Innovation and Entrepreneurship, ISSN 2192-5372, Springer, Heidelberg, Vol. 9, Iss. 1, pp. 1-23, https://doi.org/10.1186/s13731-020-00122-8

This Version is available at:
https://hdl.handle.net/10419/259602

Woods et al. Journal of Innovation and Entrepreneurship (2020) 9:13
https://doi.org/10.1186/s13731-020-00122-8

Journal of Innovation and Entrepreneurship

RESEARCH Open Access

Predicting the success of entrepreneurial campaigns in crowdfunding: a spatio-temporal approach

Clinton Woods[1], Han Yu[1*] and Hong Huang[2]

* Correspondence: han.yu@unco.edu
[1]Department of Applied Statistics and Research Methods, University of Northern Colorado, Greeley, CO 80639, USA
Full list of author information is available at the end of the article

Abstract

As an alternative to traditional venture capital investment, crowdfunding has emerged as a novel method and potentially disruptive innovation for financing a variety of new entrepreneurial ventures without standard financial intermediaries. It is still unknown to scholars and people who use crowdfunding services whether the crowdfunding efforts reinforce or contradict existing theories about the dynamics of successful entrepreneurial financing as well as the general distribution and use of crowdfunding mechanisms. This paper presents new results obtained from investigating the Kickstarter campaign data of over ninety-nine thousand projects totaling about 1 billion USD in pledges from 2009 until the most recent 2017 through dynamical spatio-temporal modeling. The funding level, the percentage of a project's goal actually raised from online communities, is used as the outcome of interest in the modeling to associate with dollar pledged and backer count that reflect the signals of underlying project quality. Evidence from the results was found to support the dynamic impact of the geographic location of a Kickstarter on its success and the associations between the observed project traits and the success of the entrepreneurial effort in the presence of the unmeasured spatio-temporal confounding. These results offer further insight into the empirical dynamics of the emerging phenomenon of online entrepreneurial financing about the role the spatio-temporal component plays in both the type of projects proposed and the association of sociocultural traits of successful fundraising with the underlying quality.

Keywords: Entrepreneurial financing, Crowdfunding, Geographic component, Spatio-temporal modeling, INLA

Introduction

One of the most critical of resources required for new ventures to succeed is financing. The financing promotes creative ideas, stimulate entrepreneurs to gather resources, hire workers, and transform resources into goods and services for society's consumption (Frank, 1998). Crowdfunding and its concepts have been around for a few hundred years as a method for raising funds for ventures that many people want or need. The idea behind crowdfunding is to gain support from a relatively large number of

small investors in order to fund a large project, thus generating the capital needed to start or maintain a venture without requiring the backing of wealthy donors. Crowdfunding is mostly viewed from the entrepreneurial perspective that startup capital is needed for the founding of a new business. Even though this is a common reason for crowdfunding, there have been other historical uses, such as war bonds to help fund a nation's military effort or helping fund construction of the base of the Statue of Liberty in New York. Overall, crowdfunding's greatest strength may be its ability to help people with an entrepreneurial spirit become business owners by overcoming the barrier that stops many: a lack of available capital, as well as its ability to transform ordinary customers into business investors (Ordanini, Miceli, Pizzetti, & Parasuraman, 2011), (Belleflamme, Lambert, & Schwienbacher, 2014), (Mollick, 2014).

With the widespread use of the internet in recent years, crowdfunding has emerged as a novel method and potentially disruptive innovation for financing a variety of new entrepreneurial ventures without standard financial intermediaries. The internet is responsible for giving people the connection they need to find investors who are willing to fund entrepreneurial efforts. There are three large companies based in the United States that have given people platforms that allow them to market their ideas to the world: Kickstarter, Indiegogo, and GoFundMe, which process billions of dollars in campaigns each year. Websites such as Kickstarter [https://www.kickstarter.com/] have made user interfaces that are easy to follow and give people an effective template when presenting their ideas, helping their users attract possible donors. The ideas presented in entrepreneurial crowdfunding cover everything from art and music to technology and food, a fact which demonstrates crowdfunding's power to open up the business world to people with any set of skills (Schwienbacher & Larralde, 2010). Crowdfunding trends of products and ideas vary in different locations, with many happening in the United States, Europe, and Australia. Crowdfunding has gained popularity over time, with an ever-increasing number of campaigns and investors participating. It was relatively popular from the start, and it has rapidly grown in prominence since then.

As an alternative to traditional venture capital investment, it is still unknown to scholars and people who use crowdfunding services what makes for a truly successful drive to obtain investors and whether the crowdfunding efforts reinforce or contradict existing theories about the dynamics of successful entrepreneurial financing as well as the general distribution and use of crowdfunding mechanisms (Agrawal, Catalini, & Goldfarb, 2010), (Burtch, Ghose, & Wattal, 2011), (Mollick, 2014). One of the most difficult parts in entrepreneurship research is dealing with sociocultural facets that have the elusive nature of preparedness, creativity, perseverance, and the capability of transforming old values into more appropriate ones as the entrepreneurial life starts (Vuong, 2016). Due to the complexity and complication arising from a diverse range of entrepreneurial goals and approaches (Schwienbacher & Larralde, 2010), a taxonomy of causes and effects in the dynamics of entrepreneurship process would be rarely complete and effective, especially when considering the sociocultural and spatio-temporal factors in the large scale.

As crowdfunding becomes more and more popular alternative financing, many researchers have explored various methods to understand the dynamics behind it. Mollick (2014) (Mollick, 2014) took a holistic view and proposed that personal networks and underlying project quality as well as geography were the most important

factors in determining the success of crowdfunding. (Mitra & Gilbert, 2014) took a different approach and focused on analyzing the language used in crowdfunding. By studying a huge corpus of texts presented in 45,000 projects, they found that phrases following certain principles such as reciprocity, scarcity, and social identity increased the chance of success. (Vuong, 2016) selected a group of factors seen as critical to the understanding of entrepreneurial efforts based on the extant literature of entrepreneurship. Evidence was found to support the relationship between sociocultural traits and entrepreneurship-related performance or traits adjusting for geographical locations.

Pinpointing the causes of all the changes and trends that appear to happen randomly at any given time or location in the ever-changing environment is something that eludes the observers of the entrepreneurial activities. If one were able to infer what the causes of a successful crowdfunding effort were at any given location and time, then this could enhance the understanding of these kinds of ventures and help future entrepreneurs make the right plans when starting a campaign. Finding the causes could also help us predict the future outcome of any given campaign or even predict the outcomes of many campaigns, thus perhaps finding a trend in its early stages before it even begins. Spatio-temporal variability is the key to understand all this.

The complexity of the underlying dependence structure of the spatio-temporal component seems to be growing visibly. Changing trends in different times and locations can be observed. Many locations have experienced a change in the types of ideas that are most successful over time and others seem to discover online crowdfunding for the first time. Understanding the dynamics of crowdfunding at many different times and locations is useful in helping entrepreneurs or small enterprises to raise the essential capital from the crowd to support their projects or business. This paper first time conducts the research exploring the spatio-temporal pattern of the success of campaigns in crowdfunding using spatio-temporal statistical approach for additional insight into the dynamics of crowdfunding.

The rest of the paper is organized as follows. We begin with the methods in "Methods" section. Specifically, "Data" section introduces the data followed by the descriptive analytics in "Descriptive patterns" section to provide insight into the past and figure out what has happened; Then the spatio-temporal model is developed for an in-depth analysis of the dynamics of a successful crowdfunding campaign in "Spatio-temporal model" section. After that, the results are presented in "Results" section. Discussion and conclusion are finally given in "Discussion" and "Conclusion".

Methods

Data

The data scraped in its original form consists of 104 csv data sets spanning the time from 2009 until 2017. When merged, the data sets contain 36 million observations, many being duplicates due to web scraping taking place each month and projects going from inception to deadline within months as well. Some may have been projects that were restated as well. The campaign IDs of the observations were used to remove the duplicates. Many observations were also missing. Therefore, variables with observations too incomplete were removed from the overall data set for this study. As a result, a data set with 99,036 observations totaling $1,064,392,179 USD in pledges was created. The

data has observations from all over the world, with most originating from the United States and North America. This data set should be a good representation of Kickstarter campaigns and possibly any crowdfunding platform that can be used on the internet to help entrepreneurs gain startup capital.

Our statistical goals are to visualize, summarize, and infer the dynamical behavior of crowdfunding. The complex spatio-temporal data are a window to the underlying complex dynamics of crowdfunding, from which the extraction and description of information are challenging. For analytical understanding of crowdfunding, the descriptive analytics via summarization and visualization are considered in conjunction with spatio-temporal modeling because they suggest relationships that can be incorporated into spatio-temporal models for the purpose of inference. The development of exploratory and diagnostic methods for spatio-temporal data is an important research topic and will remain so in the future.

Descriptive patterns

Prior to any statistical modeling for the dynamics underlying the complex data set, the variability of the data across locations over time and major variables appeared in the literature are described beginning with the information on worldwide campaigns. The goal of this part is to develop initial evidence about the nature of crowdfunding, which is appropriate for an evolving topic in the evolving field of entrepreneurship. Figure 1 shows a general map of the geographical distribution of the locations of the Kickstarter campaigns. Notice that places that speak English tend to use the Kickstarter platform at much higher rates than other nations around the world. Table 1 gives a solid look at which nations are using this platform the most. Notice that the US makes up a majority of the campaigns while many of the other nations make up just enough observations to possibly produce some kind of understanding of how a successful Kickstarter happens across countries.

The first bar plot in Fig. 2 shows the change of the status (canceled, failed, live, successful, and suspended) and the associated number of campaigns for each status across the cities observed in the USA superimposed with a median pledge in USD. Some cities appear to be more popular than others for a campaign, while there is variability when looking at successful cities versus unsuccessful cities. A chi-squared test against the null hypothesis that status and city are independent reports the observed chi-squared statistic x^2 of 19,590 with P value < 0.0001 based on 12,840 degrees of freedom, indicating that status is related to location. The influence of geographic location will be considered for incorporation into the model for statistical inference. The second bar plot in Fig. 3 shows the status based on the categories with median pledge USD by line: canceled, failed, live, successful, and suspended. There is a noticeable variation in popularity across the categories, while different categories also show different levels of success. A chi-squared test against the null hypothesis that status and categories are independent reports a P value < 0.0001 based on the observed x^2 of 22,253 with degrees of freedom 56. This indicates that the status of a campaign associated with the categories of the campaign. In both bar plots, notice that some categories and geographic locations seem much more successful than their counterparts. When looking into the live median pledges per city in the data, some cities are showing a higher rate of pledges

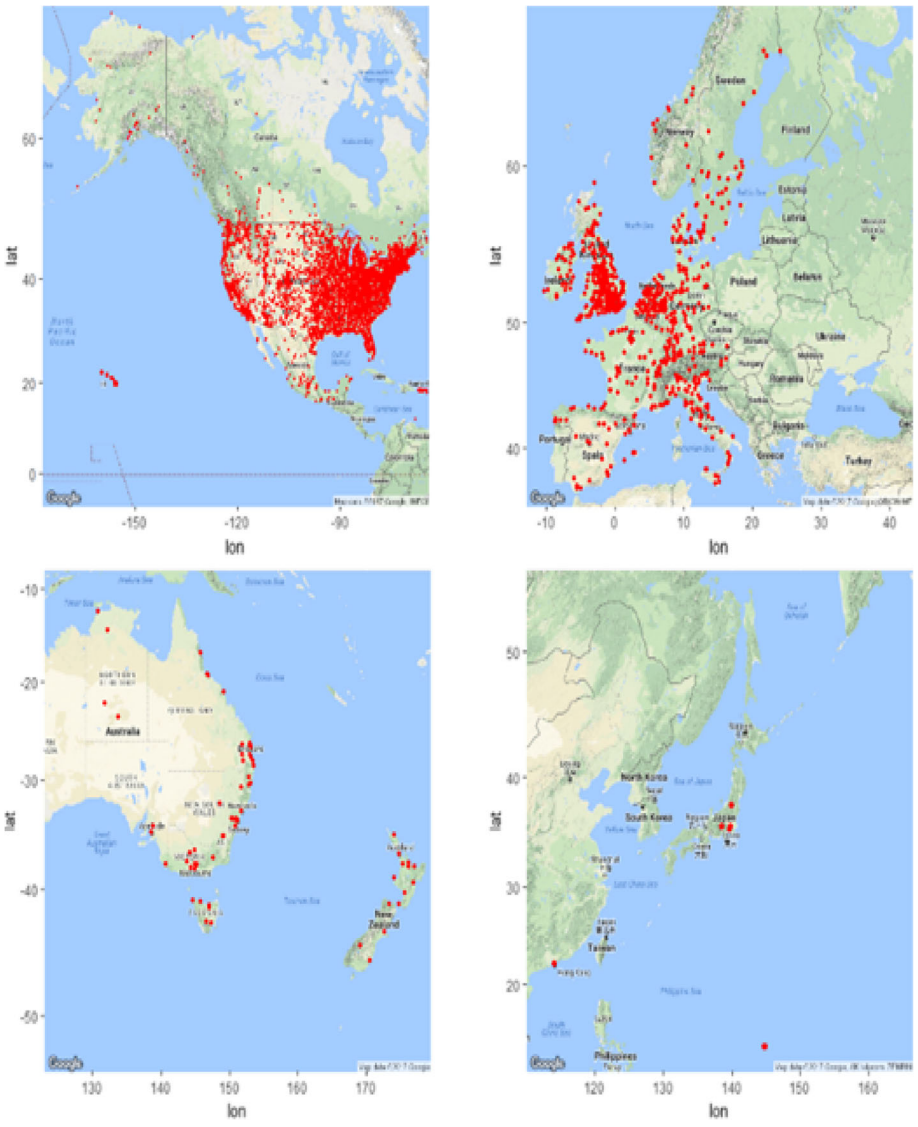

Fig. 1 Worldwide distribution of Kickstarters

happening at the current time than the historic popularity of the city would have you believe. A chi-squared test against the null hypothesis that categories were independent of the US state where they are reports a P value < 0.00001 for the observed x^2 of 7043 on 686 degrees of freedom, indicating that the categories are dependent on what US state the campaign is in. This suggests possible trends developing for both the cities and categories of Kickstarter campaigns.

The histogram in Fig. 4 shows goals in red versus the actual pledges by donors in blue. Notice that the goals seem to ask for more money than what the actual campaigns seem to achieve. Pledge goals seem to lay in a narrower region than the more variable pledged money. The Kickstarters overall seem to overestimate and set goals much higher than they will actually receive. Many in the red zone of the histogram may fail, while the campaigns existing in the purple zones may represent successes. Based on the fact that both can be log-transformed to a normal distribution, a t test against the null

Table 1 Descriptive statistics of international distributions

Country code	Mean campaign USD	Projects per country	Mean backers	Total pledged	Total backers	Mean days making
USA	11043.51	94692	119	1045732426	11268511	47
GBR	5017.59	1582	70.58	7937836.51	111662	41
CAN	6060.78	755	96.63	4575895.32	72953	43
AUS	6486.47	362	90.76	2348103.52	32854	53
DEU	10139.95	280	96.7	2320574.92	27077	44
FRA	8816.48	189	106.26	1398167.5	20084	53
SWE	15856.03	75	213.49	1189202.75	16012	46
NLD	13359.25	138	112.93	792080.14	15584	59
ESP	5806.3	138	80.54	523976.43	11115	23
SGP	6898.2	86	107.19	593245.97	9218	33
ITA	3802.86	177	38.84	1503632.3	6874	41
DNK	6157.57	76	87.39	467975.33	6642	41
NZL	4842.68	79	62.3	382572.16	4922	48
MEX	1099.21	184	20.98	202255.61	3860	19
CHE	11747.62	53	69.45	622623.98	3681	48
NOR	7860.21	43	57.53	337989.4	2474	70
IRL	3369.78	55	42.64	135632.81	2345	45
BEL	4325.63	42	49	225905.79	2058	47
AUT	4992.15	28	35.5	62422.46	994	35
JPN	2052.52	1	59	2052.52	59	9
LUX	0	1	0	0	0	34
Total	6596.25	99036	76.98	1071364179	11618979	42

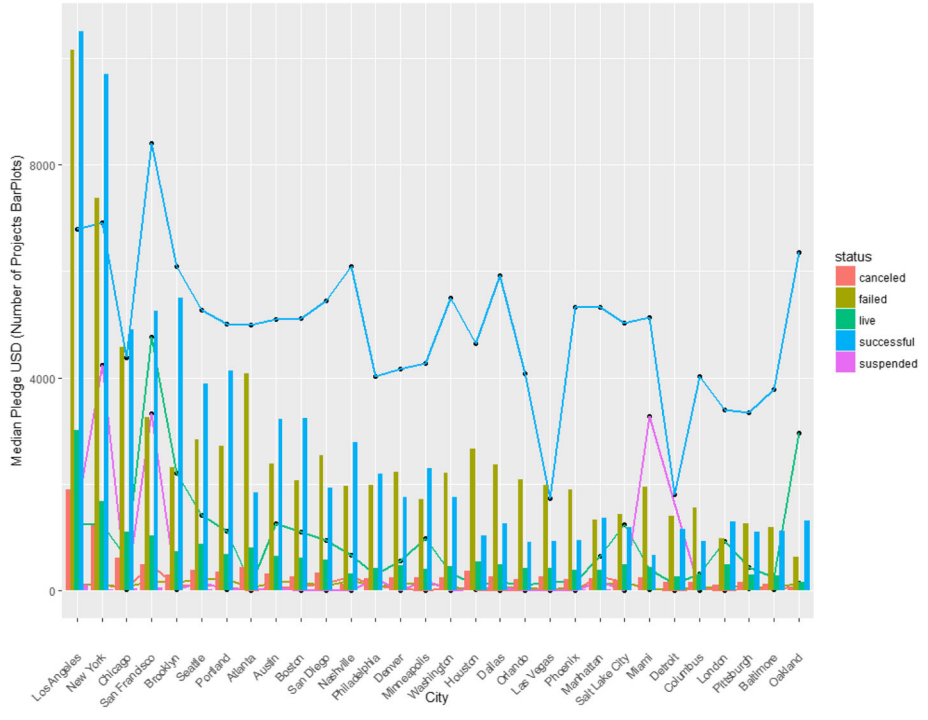

Fig. 2 Bar plot of the number of projects with line representing the median pledge in USD

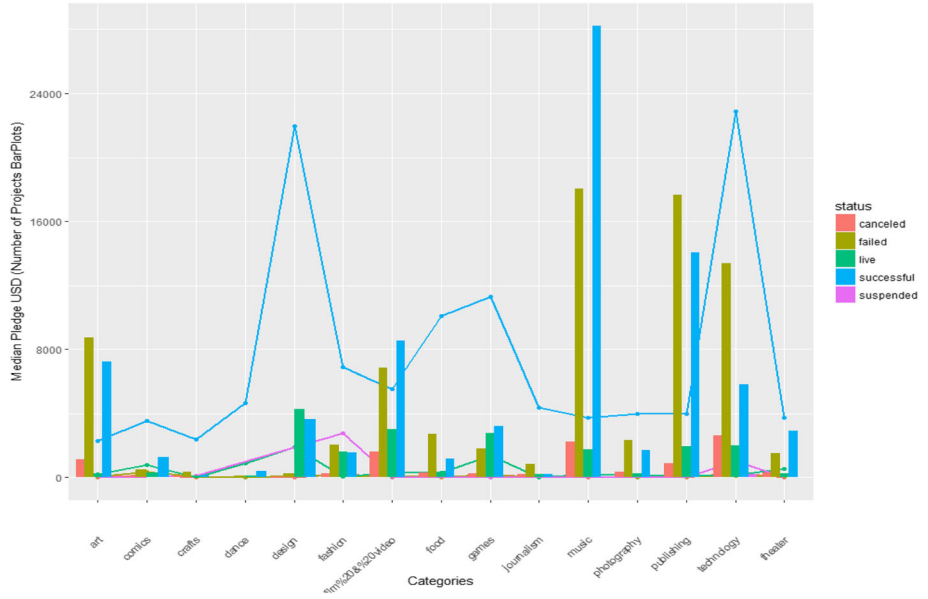

Fig. 3 Bar plot of the number of projects per category with line showing the median pledge USD

hypothesis that the goal USD is higher than the pledged USD actually obtained reports a P value < 0.0001.

The histogram in Fig. 5 displays the percentage of the goal that a project obtains by its end on a log scale due to large right skewness. There are two modes for the failed and live categories sitting near a low percentage of backers and for the success category. On one hand, many of the failed projects do not even come close to reaching their goals. On the other hand, most successful projects got backed only by making the

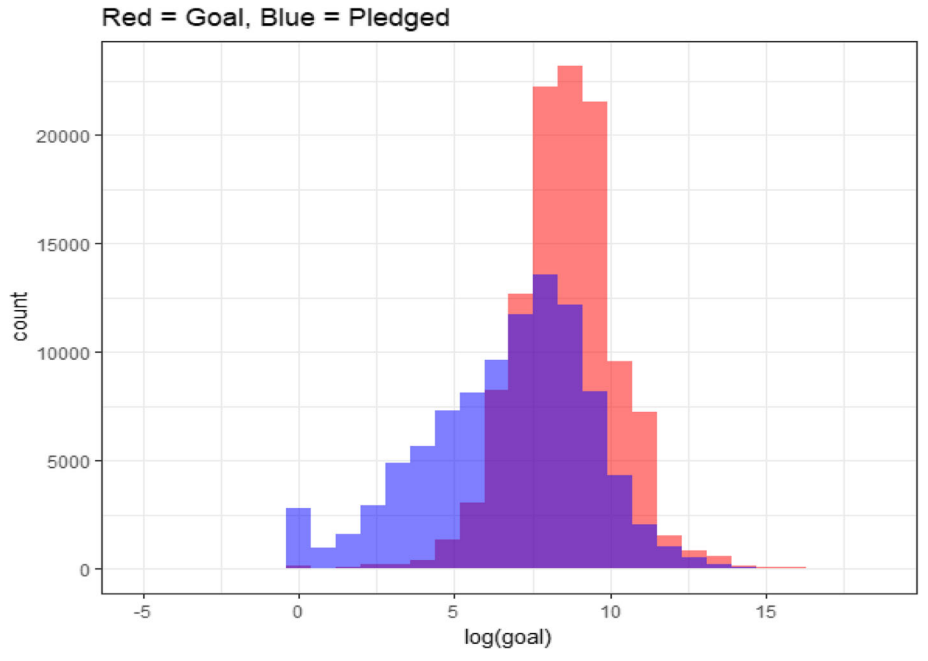

Fig. 4 Histograms of goal (red) versus pledge (blue)

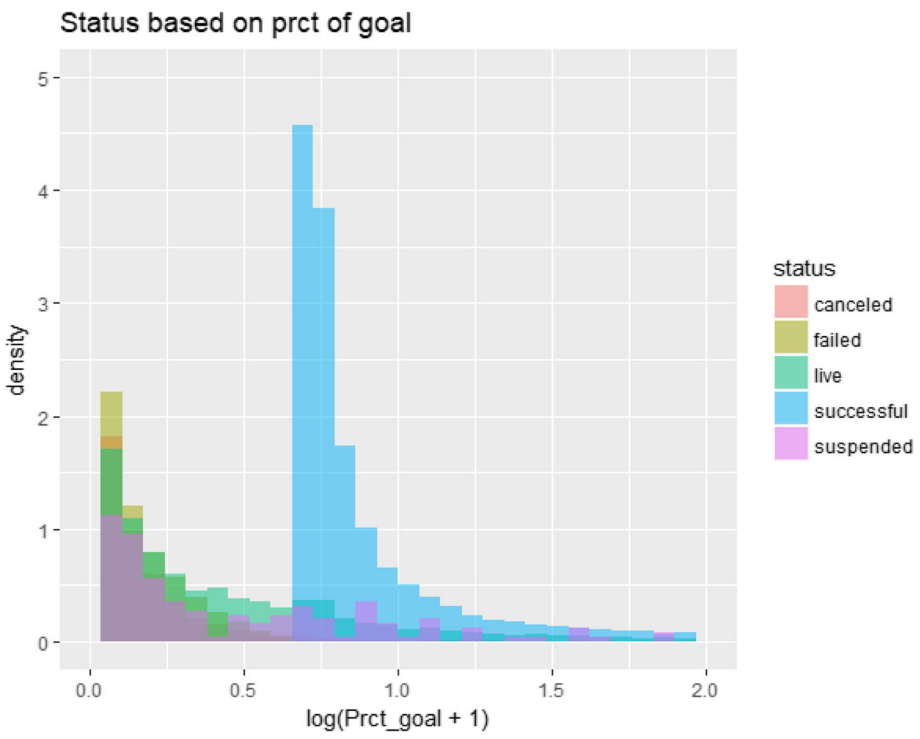

Fig. 5 Histograms of percent goal based on categories

minimum funding to reach their goals, while some others got much more than a 100% backing for their projects. This brings up the question of how a campaign achieves more than the goal set. These represent a further possible group of extremely successful campaigns that go above and beyond just a simple success.

Figure 6 displays a histogram of the number of backers across status with blue being successes and brown failures. There are others that have numbers too small to notice, such as the suspended category. Notice that there are many more backers on the more successful projects, whereas the failed projects seem to stay down near the lower end of the backer count. This shows a correlation between project success and backer count in which projects with larger numbers of backers are more likely to be successfully funded. Also notice that failed and successful campaigns cross over in the number of backers with some failures having a lot of backers and some successes having very few.

The maps in Fig. 7 exhibit the average number of backers in each US state across all the years being studied, beginning with 2009 in the top left and concluding with 2017 in the bottom right. You can see that there are more backers every year, indicating the increasing popularity of this platform to make money for a project. As the overall popularity of the projects goes, they have gone from a mean of 40.62 backers per project in 2009 to a mean of 154 backers per campaign in 2016, with 2016 being the last year for which the data includes for all 12 months. Each state's backer count increases differently when compared to its neighbors as overall popularity is growing. This seems to display local trends happening in states in different years. Montana's projects in 2017, on average, seem to be very popular compared with many other states in the US. A chi-squared test against the null hypothesis of independence between mean backers of states versus years reports a P value < 0.00001. In contrast to the state maps in Fig. 7

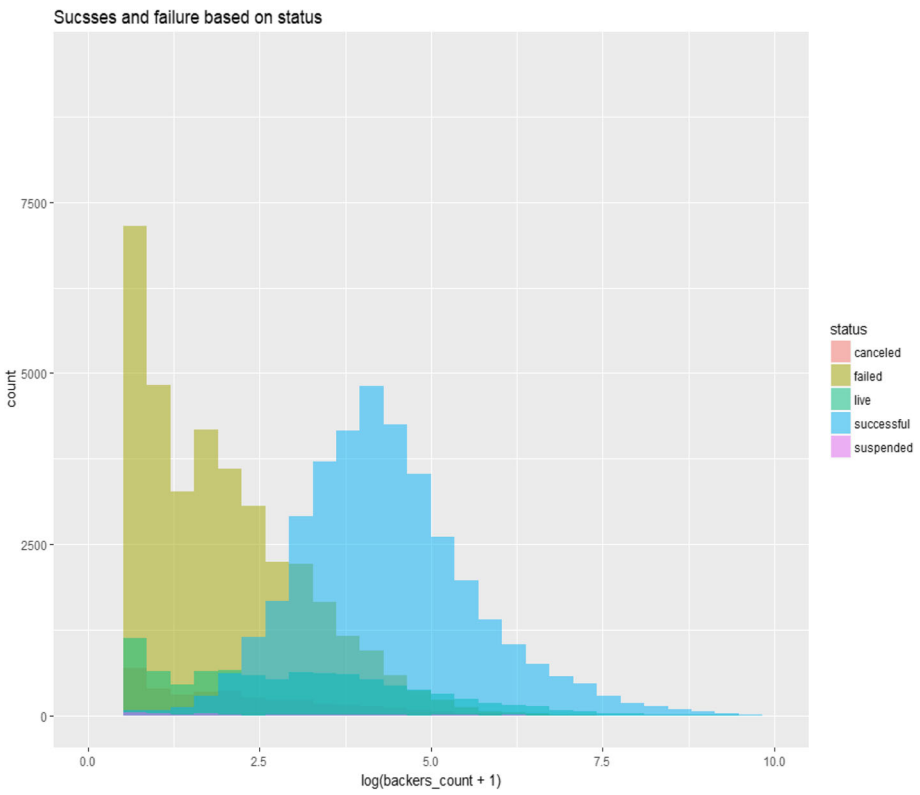

Fig. 6 Histograms of backers count based on categories

that show the variability of the number of backers over time, Fig. 8 displays the variability in mean US dollar amount of pledges per county in the United States. The darker the color, the higher the amount donated from that county. There exist areas in which campaigns are gaining quite a bit of money from pledges compared with other parts of the US. This map seems to show that many of the campaigns have concentrations of entrepreneurship where people come up with successful ideas. As seen in the bar plot in Fig. 2 provided earlier, there are certain cities that see a larger portion of the Kickstarter activity, which could also be the areas that pull in the most money on average. Another set of maps that are too cluttered to add to this paper show the mean US dollar amount of pledges per county in the United States from the years 2009 to 2017. The maps seem to show that more and more counties are donating more money as time goes on. The states and counties all appear to sit around more populated areas, which could be attributed to population density or other latent variables that explain why smaller areas have a more meager mean spending on Kickstarter. Due to this being an internet platform, one may think that the mean money made per project in each county would be more similar, but this is not the case. The descriptive statistics for US states are presented in Table 2, showing how the states differ numerically.

The bar graph in Fig. 9 represents the percentage of backers in city populations across all time. You can see the change in how popularity is viewed: for example, Los Angeles being the city with the most overall backers might only be so because of a larger population. San Francisco and Salt Lake seem to have had an extremely successful number of backers based on population size. So even in an extreme case such as Salt

Fig. 7 Maps of observed mean backers per state over the years

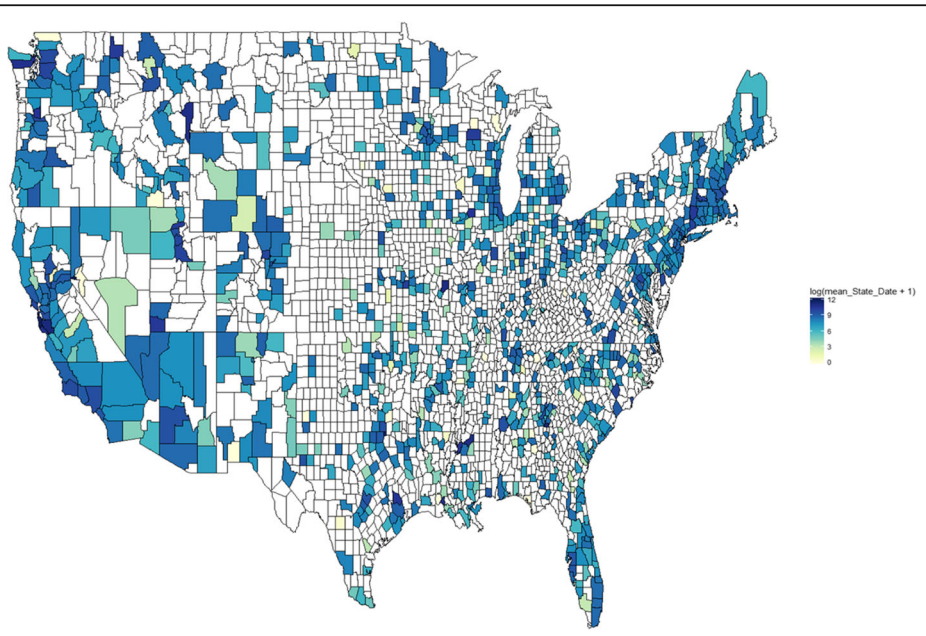

Fig. 8 Map of county map showing the mean pledges USD per county

Table 2 US states descriptive statistics

State	Mean campaign USD	Projects per state	Mean bakers	Total pledged	Total backers	Mean days building	Median percent of goal
California	21572.54	15694	215.5	338559439	3382025	52	31.3
Non-USA	9986.81	13095	110.3	130777355	1444720	46	19.1
New York	11667.81	9701	125.2	113189460	1214094	49	100
Texas	8348.34	6288	104.7	52494414	658492	47	9.4
Illinois	11628.06	3873	128	45035495	495840	51	25.7
Oregon	13294.05	2684	139.6	35681238	374583	45	55.9
Massachusetts	12059.9	2897	130.1	34937543	377002	50	55.7
Utah	19153.32	1641	217	31430608	356156	57	30.3
Georgia	10058.05	2336	90.9	25507227	230716	40	4
Colorado	10093.5	2527	142.5	25506277	360046	46	17.4
Arizona	10916.01	2166	104.6	23644468	226552	45	9.5
Florida	4259.6	3810	49.2	16229080	187629	43	3.5
Tennessee	7050.33	2272	79.9	16018366	181552	41	33.1
North Carolina	5720.82	2718	72.8	15549209	197948	43	9.4
Pennsylvania	5940.57	2527	66.1	15011844	166949	45	21.9
Minnesota	8233.6	1733	89.7	14268831	155512	46	41.9
Ohio	4968.83	2736	64.8	13594736	177344	45	8.7
Washington	10408.27	1065	129.1	11084816	137458	49	20.3
Michigan	5856.28	1606	66.5	9405196	106796	43	14.1
Nevada	7168.45	1203	64.9	8623655	78029	39	4
Washington DC	6968.68	1171	85.1	8160327	99672	42	17.2
Louisiana	8304.32	981	101.7	8146541	99772	49	13
Maryland	6458.89	1241	65.4	8015491	81156	49	13.9
Virginia	4749.22	1672	62.6	7940706	104651	45	11.1
New Jersey	5793.69	1228	67.5	7114655	82942	48	6.2
Connecticut	8475.96	738	60.1	6255260	44334	53	15.8
Wisconsin	5046.63	1188	84.2	5995398	99998	52	19.6
Indiana	5282.96	1131	64.9	5975035	73394	42	5.7
Montana	20556	284	175.6	5837903	49881	52	53.9
Kansas	9064.92	431	50.4	3906984	21725	43	8.7
Missouri	4049.56	773	47.7	3130311	36851	40	10.2
Rhode Island	8755.71	341	91.9	2985697	31365	42	59.6
New Mexico	5507.99	426	72.2	2346406	30741	47	12.4
South Carolina	3858.75	555	41.9	2141610	23271	37	5
Oklahoma	2969.17	693	40.5	2057636	28102	40	10.6
Nebraska	5133.82	357	84.5	1832776	30162	46	6.6
Alabama	4089.05	440	61.2	1799183	26932	44	3.6
Kentucky	3839.89	406	62.7	1558996	25454	42	6.9
Iowa	3533.7	428	63.6	1512424	27222	41	12.6
Arkansas	3930.71	372	46.3	1462225	17231	35	3.9
Mississippi	5901.73	226	65.6	1333793	14818	44	5.6

Table 2 US states descriptive statistics *(Continued)*

State	Mean campaign USD	Projects per state	Mean bakers	Total pledged	Total backers	Mean days building	Median percent of goal
Idaho	5030.64	256	48.2	1287845	12347	45	11.5
Delaware	15272.89	59	139.4	901100	8228	73	21
New Hampshire	4296.77	185	59.2	794903	10955	44	17.5
Vermont	4578.18	120	59.8	549382	7176	50	104.6
Wyoming	5148.39	93	54.4	478800	5058	59	10.1
West Virginia	2943.4	133	41.7	391472	5550	33	5
Maine	3469.31	108	39.1	374685	4221	50	9.8
South Dakota	2365	122	28.4	288531	3466	39	4.9
North Dakota	2253.05	106	26.9	238824	2861	39	12.5

Lake having a very small comparative population, it has quite a few project backers. This also seems to be the case when looking at the money each city made over the whole time period. San Francisco still seems to reign supreme as the city that has made the most money in a Kickstarter. This appears true even when Los Angeles has the most apparent projects based on both successes and overall volume. What truly drives smaller cities to be more successful per capita than larger cities does indeed deserve a deeper analysis to understand what makes a successful spatial location other than the largest population.

Figure 10 is a word cloud based on the most common words used in the titles of successful campaigns, where the larger the word the more common its usage. The words that were found equally commonly in successful and unsuccessful campaigns are removed so as to focus on the words that appear more often in successful campaigns. In the word cloud, you can see the words “record,” “debut,” and “tour” as top words, which are words you may think of as relating to music as shown in Fig. 3, where music is the most common category and has the most overall success. There are also many other words that appear in the more successful campaigns. Blurbs, which are small

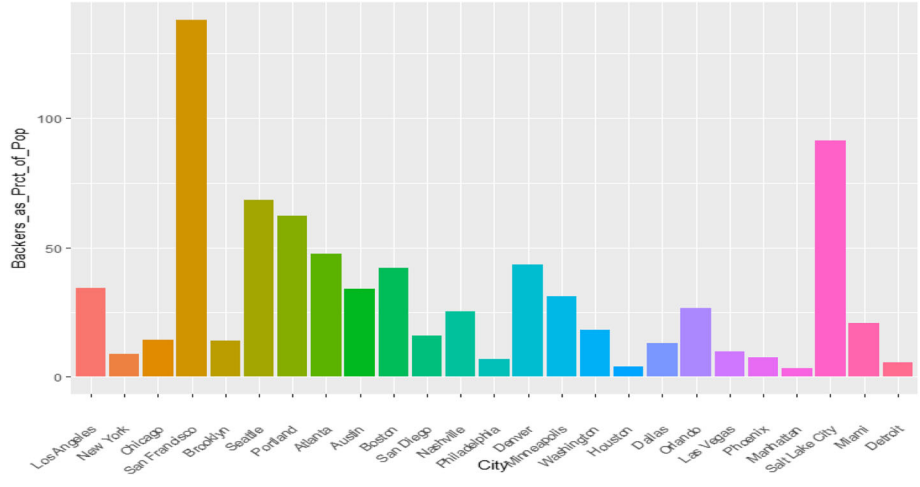

Fig. 9 Backers as a percentage of city population all time

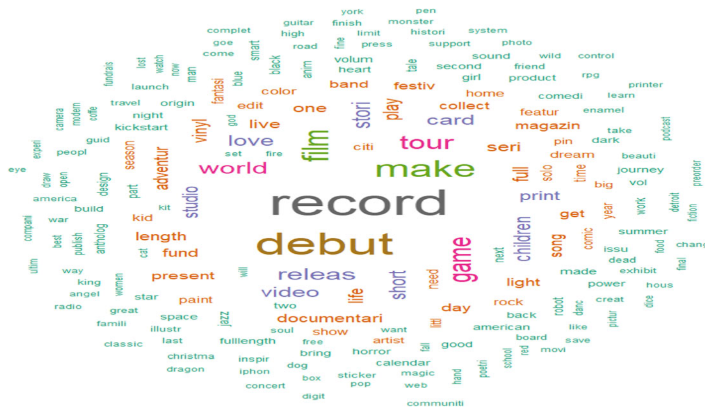

Fig. 10 Word cloud showing the most popular words in a successful project

paragraphs explaining the Kickstarter campaigns, might also be found within the data sets. Often times, these blurbs provide similar words when analyzing successful campaigns. Blurbs may help refine what makes a successful campaign further than the numerical data alone.

The blurbs and titles consist of unique sentences for the campaign that they are attached to. A factor was then created that was weighted by attributes hidden in the blurbs and titles. To make the factor, the main data set was separated into two unique data subsets. The first weighting for the factor was created by taking the two subsets of the data, one being all failed campaigns and the other all successful campaigns. The tm, snowballC, and wordcloud packages in R were used to prepare and separate the words used in blurbs and titles based on being a failed or successful campaign. After the data was prepared, a count was then taken of the words that appeared the most in both successful and failed campaigns. This was then further cleaned by taking the top 500 most common words of both successful and failed campaigns. The last step was then to only take the top words that uniquely appeared on successful campaigns, thus removing any words that appeared in both of the top 500 data sets. For this weighting, the top 50 unique words left in both blurbs and titles were extracted.

To get the other weighting for the factor, a new subset of the data was made. This data set consisted of only successful campaigns that were in the 90th percentile in terms of the number of backers or of dollars earned. The tm, snowballC, and wordcloud packages in R were then used to prepare and separate words used for blurbs and titles. Words that were related to extremely successful campaigns were often not unique, suggesting similar marketing strategies. A count was taken of the most common words in this new category. The top 50 most common words for both blurbs and titles were extracted for weighting. This gave us two categories for blurbs and two categories for titles, a total of four new sets of words. For each of the new sets of words, the grepl function in base R was used; this function queries strings of words for specific sets of words or characters and returns a TRUE or FALSE value. After using grepl for each of the new sets of words, four new columns were created that contained either TRUE or FALSE based on if a title or blurb had one of the words in the four sets. The TRUE and FALSE values were changed to 1 representing TRUE and 0 representing FALSE. The new ones and zeros were used as weights for summation. After the four

columns were summed, they gave a new factor column that ended up being called Twords with levels coded as 0, 1, 2, 3, and 4 as shown in Table 3.

In addition to the examined variables here that are major variables having appeared in the literature, there may be many more relevant variables that need to be identified and screened to predict the behavior of a project. The semiparametric structured modeling is natural for complex data with variables of different types in different formats and technical variables of high dimensions. The spatio-temporal data consists of strings of words, such as the title and what is written about the campaign to help entice donors to give their project a shot as well as many relevant covariates important to the possible trends and evolving behavior of crowdfunding projects.

Spatio-temporal model

More recently, standard machine learning algorithms are the only major methods applied to study the elements of a successful crowdfunding campaign in the existing literature. As is well known, most standard machine learning algorithms are based on the independence of the observed values replicated from the same model, where these observed values are considered independent realizations of the same random variable. The First Law of Geography states: "Everything is related to everything else, but near things [in space and time] are more related than distant things" (Tobler (Tobler, 1970)). When the observed values are anchored in space and time, the assumption of independence is no longer realistic. The dependence structure of crowdfunding campaigns rising from space and time is the key to understand the dynamics of crowdfunding and should be considered in the modeling.

To mitigate unobserved spatial confounding, evaluate the impact of the geographic location of a Kickstarter, and further understand the dynamics of online entrepreneurial crowdfunding efforts (Mollick (Mollick, 2014), Vuong (Vuong, 2016)), we employ a hierarchical dynamical spatio-temporal framework (Cressie and Wikle (Cressie & Wikle, 2011)) to develop a hurdle model with the data from the Kickstarter platform. The dynamical hierarchical hurdle model describes the variability of the outcomes that are more correlated when close in space and time than outcomes that are collected further away. The hurdle model was employed to predict the funding level, the percentage of a project's goal actually raised from online communities, with the dollar pledged and backer count of a crowdfunding effort that reflects the signals of underlying project quality in conjunction with spatio-temporal component. The predicted funding level can readily be shifted to the outcome of success and failure.

The data can be considered a realization of a dynamical process for funding level indexed by geographical locations and time points in a study region D (i.e., USA in this

Table 3 Statistics of the new Twords variable

Factor/stat	0	1	2	3	4
Mean backers count	72.6	113.3	153	180.7	241.9
Median backers count	10	13	18	26	33
Mean pledged USD	6536.1	10036.8	14380.5	19797.2	22708.2
Median pledged USD	475	660	1000	1521	1790
Factor obs count	24172	44233	25511	4711	409

paper) that resides in the 2-dimensional space R^2 (US states) and the 1-dimensional temporal line R (years)

$$Y(s,t) \equiv \left\{ y(s,t), (s,t) \in D \subset R^2 \times R \right\},$$

where s is the geographical location in the study region D of the United States and t is the year from 2009 to 2017. The spatial component of the data was modeled on US states.

Spatio-temporal components

The predictor η_{ist} for the spatio-temporal components of a Kickstarter project is represented as follows,

$$\eta_{\text{ist}} = b_0 + b_1 x_i + u_s + v_s + \gamma_t + \phi_t + \delta_{st}, \tag{1}$$

where b_0 is the intercept, b_1 is the vector of linear fixed effects of the vector of observed covariates x_i, u_s is spatially structured effect; v_s is the spatially unstructured effect, γ_t is the temporal process, $\varnothing_t$ is independent temporal effect, and δ_{st} is the spatio-temporal interaction.

The structured spatial effect

$$u_s | u_{s'}, s' \neq s, \tau_u \sim N\left(\frac{1}{n_s} \sum_{s' \sim s} u_{s'}, \frac{1}{n_s \tau_u} \right), \tag{2}$$

where n_s is the number of the neighbors of state s, $s \sim s'$ indicates that the two states s and s' are neighbors, and the precision parameter τ_u is represented as $\theta_1 = ln(\tau_u)$ and the prior is defined on θ_1. v_s is the spatially unstructured effect.

The structured temporal effect of the component can be represented as the random walk process of order two (RW2). The random walk process of order two for the Gaussian vector $\gamma = (\gamma_1, \ldots, \gamma_T)$ is constructed assuming independent second-order increments.

$$\gamma_t | \gamma_{t-1}, \gamma_{t-2} \sim N(2\gamma_{t-1} - \gamma_{t-2}, \tau_\gamma^{-1}). \tag{3}$$

The precision parameter τ_γ is represented as $\tau_\gamma = e^{\theta_2}$ with a prior on θ_2. The unstructured component $\varnothing_t$ of the model is represented with an interchangeable model. This model simply defines $\varnothing = (\phi_1, \ldots, \phi_T)$ to be a vector of independent Gaussian random variables with mean zero and precision $\tau_\varnothing$, i.e.,

$$\phi_t \sim N\left(0, \frac{1}{s_i \tau_\phi} \right), \tag{4}$$

where $s_i > 0$ is a scalar.

The δ_{st} is the spatio-temporal interaction between the spatially and temporally structured effects u_s and γ_t called type IV interaction by Blangiardo and Cameletti (2015) (Blangiardo & Cameletti, 2015), which can be represented by a structured matrix R_δ of rank $(n-1)(T-2)$ for a RW2. The structured matrix can be written as a Kronecker product of $R_\delta = R_u \otimes R_\gamma$. This assumes that the temporal dependency structure for each area is not independent from other areas and areas depend on the temporal patterns of

their neighbors as well. This type of interaction is the most appropriate for the data under the assumption that Kickstarter is online assuming so that interaction is highly dependent.

Two linked hurdle mediation models

There are 11,729 campaigns out of 99,036 that never received any backers at all, resulting in these 11,729 campaigns having zero USD pledged. There are two typical models available to account for this potentially high occurrence of zeros in failed campaigns (Hu et al. 2011). The first is zero-inflated models assuming that all projects have a certain chance to obtain a zero, i.e., all zeros have two different origins: structural origin and sampling origin. In such a model, sampling zeros occur by chance, while other zeros are observed due to some specific structure in the data. The second is a hurdle model with a latent factor Z_i taking 0 and 1 with 1 indicating an observation that passes a hurdle (success) and is defined as a positive count and 0 indicating an observation that does not pass the hurdle (failure) and is defined as a zero count. The assumption for the hurdle model is that all zero observations are considered from one structural source. There are a considerable number of failures obtaining zero count of backers, it is plausible to consider a campaign with zero backers to be a complete failure representing an anomaly in contrast with the rest of the failures that gained at least a positive count of backers, i.e., the zero counts most likely are from one structural source and need to be studied separately while the positive counts of backers have a sampling origin, implying choice of the hurdle model over the zero-inflated model. The binary latent variable Z_i represents the origin of data, with 1 representing a positive count and 0 a zero count for the hurdle model

$$Z_i = \begin{cases} 0, & with\ probability\ \ 1-\pi_0 \\ 1, & with\ probability\ \ \pi_0 \end{cases} \tag{5}$$

where π_0=Pr(Z_i = 1) and logit(π_0) = η_{ist}. Conditional on the binary latent variable Z_i, the hurdle models for $X_{\text{ist}}^{(1)}$ representing backer count and $X_{\text{ist}}^{(2)}$ USD pledged for project i can be specified as the finite mixture models

$$p\left(x_{\text{ist}}^{(j)} | Z_i = z_i\right) = \begin{cases} 1-\pi_0, & z_i = 0 \\ \pi_0 f\left(x_{ist}^{(j)}\right), & z_i = 1 \end{cases} \quad j = 1, 2 \tag{6}$$

where $f(x_{\text{ist}}^{(j)})$ is the probability density function for the positive count of $X_{\text{ist}}^{(j)}$. Backer count $X_{\text{ist}}^{(1)}$ has many campaigns with a smaller number of backers and a very large and gradual tail of backers. Gamma distribution with predictor η_{ist} makes the most sense for $X_{\text{ist}}^{(1)}$ due to the heavy right skew of the data, i.e., $X_{\text{ist}}^{(1)} \sim$Gamma($s\vartheta$, μ_{ist}), where s is a fixed scaling factor. The ϑ is reparameterized as $\vartheta = e^{\theta_3}$ *with* the prior log-Gamma for θ_3.

The pledged USD $X_{\text{ist}}^{(2)}$ can be specified as lognormal distribution, i.e., log($X_{\text{ist}}^{(2)}$) ~ N(η_{ist}, τ_1) by transformation using the max of the observations subtracted by the values with the log-gamma prior for θ_3 in the reparameterization $\tau_1 = e^{\theta_4}$.

Funding level model

The funding level, the percentage of a project's goal actually raised from online communities, is used as the outcome of interest in the modeling to associate with dollar pledged and backer count that reflect the signals of underlying project quality. The percent of a project's goal actually raised decides if a campaign is a failure or success since only campaigns with 100% funded or more will be deemed successful by the Kickstarter platform. The spatio-temporal component can play an important role in both the type of projects proposed and the sociocultural traits of successful fundraising related to the underlying quality. Twords, city population, length of a Kickstarter, length of time for preparation, categories such as music or technology that a project falls in, US county, US states, and years that a campaign takes place are used as the predictors for additional covariate information on each effort. All covariates were included based on the availability of the information for the Kickstarter.

The funding level Y_{ist} is specified as a lognormal distribution

$$\log(Y_{\text{ist}})|l_{\text{ist}}, \tau_2 \sim N\left(l_{\text{ist}}, \tau_2^{-1}\right) \tag{7}$$

conditional on the predictor consisting of the two model components of pledged USD and backer count

$$l_{\text{ist}} = \beta_0 + g_1\left(X_{\text{ist}}^{(1)}\right) + g_2\left(X_{\text{ist}}^{(2)}\right) + \beta x_i + u_s + v_s + \gamma_t + \phi_t + \delta_{st}. \tag{8}$$

The log-Gamma prior was used for θ_5 in $\tau_2 = e^{\theta_5}$.

As the hierarchical model is richly parameterized to deal with the large spatio-temporal data sets, known as the "big n problem" (see Banerjee et al. (Banerjee, Carlin, & Gelfand, 2004), Page 387; (Lasinio et al. 2013)), and the functional form of the posterior distribution is nonstandard and unknown in practice, simulation-based MCMC is not computationally feasible for the Bayesian inference. Integrated Nested Laplace Approximation (INLA) proposed by Rue et al. (Rue & Held, 2005; Rue, Martino, & Chopin, 2009) was employed as a deterministic algorithm to perform approximate fully Bayesian inference as a valid and computationally effective alternative to the simulation-based Monte Carlo Markov chain (MCMC) method. INLA was developed based on the Laplace method of transformation to approximate the integrand with a second-order Taylor expansion around the mode and computes the integral analytically. It provides faster and more accurate results in shorter computing time compared with the MCMC scheme, especially for latent Gaussian models with large-scale data (Rue et al. (Rue et al., 2017), Bivand et al. (Bivand et al., 2015) and Ferkingstad et al. (Ferkingstad et al. 2017)). Bayesian hierarchical modeling with latent Gaussian processes has proven very flexible in capturing complex stochastic behavior in hierarchical structures in high-dimensional spatial and spatio-temporal data (Opitz (Opitz, 2017)).

We trained the model with the strategy of cross-fitting (Chernozhukov et al. (Chernozhukov et al., 2018)), which provides an efficient form of data-splitting into four samples of equal size so that the model components can be trained and tested at each step in the proposed multi-stage fitting procedure. The first subset of the data was used to train the hurdle model component. The second subset of the data was used to test the hurdle model component and then can be used to predict the backer count and the pledged USD after the hurdle model component is fitted. The third subset of the data

was used for testing the accuracy of the predicted backer count and pledged USD and training the funding level model using the two predictions. The fourth subset of the data will be used to evaluate predictive ability using correlation, which gives an evaluation of how accurate this method is at finding a successful Kickstarter campaign. All the components will be trained based on the assumption that they are both spatially and temporally dependent.

The first random subset data was used to train the model for Z_i. The deviance information criterion (DIC) was used for model fit. The lowest DIC was 15649. After predicting zeros and ones for each Kickstarter campaign on the second subset data, the projects that are predicted as Z_i=1 from the second subset data will then be picked out to be used to train subsequent backer account and pledged USD components.

The backer count component based on the hurdle model is used for further investigating traits attributing to the success of a campaign in the Kickstarter campaigns obtaining positive counts of backers. The model was trained using the second subset data and tested on the third subset data. The model predicted 42% of the actual values. The best model is based on the smallest DIC = 65,898.29 with interaction IV component compared with any other model without interaction IV component, which has a DIC greater than 160,000.

The trained model for pledged USD was able to predict 41% of the actual values in the third subset data. The best fit model for the pledged USD with a DIC of 83,180.84 compared with any other model with a DIC larger than 150,000 that did not apply the interaction IV of the spatial and temporal components.

The predicted $\hat{X}_{\text{ist}}^{(1)}$ and $\hat{X}_{\text{ist}}^{(2)}$ can now be added to the third subset data for the funding level model, each of which gives more than 40% of the information that could have been obtained if getting backer count and pledged USD was possible before one launches a campaign. The third subset data with the newly predicted variables was used in the funding level model. After the funding level model was trained, the fourth subset of the data was used as a test data set.

Results

The model with type IV interaction has a DIC of 90,578.29 and obtains the best fit compared with other models trained of DIC's near 200,000. Figure 11 shows the scatterplot separated by two red lines, which represent 100% of the actual (horizontal) and predicted (vertical) goal. Points in the top right section or the bottom left section represent accurately predicted campaigns as a successful campaign or a failed campaign respectively. This spatio-temporal model with type IV interaction was able to predict successes and failures 79% of the time regardless of how close to the actual value the prediction is. Being that the main goal of most entrepreneurs is to have a successful campaign, predicting how successful over 100% is just an additional beneficial piece of information. Different states do have better successes than others.

Figure 12 shows the posterior marginal distributions for the model hyperparameters (top and middle rows of panels) and the posterior means pattern with 95% pointwise posterior intervals for all states (bottom panel). The concentrations of the posterior marginals of the model components are all significantly different from

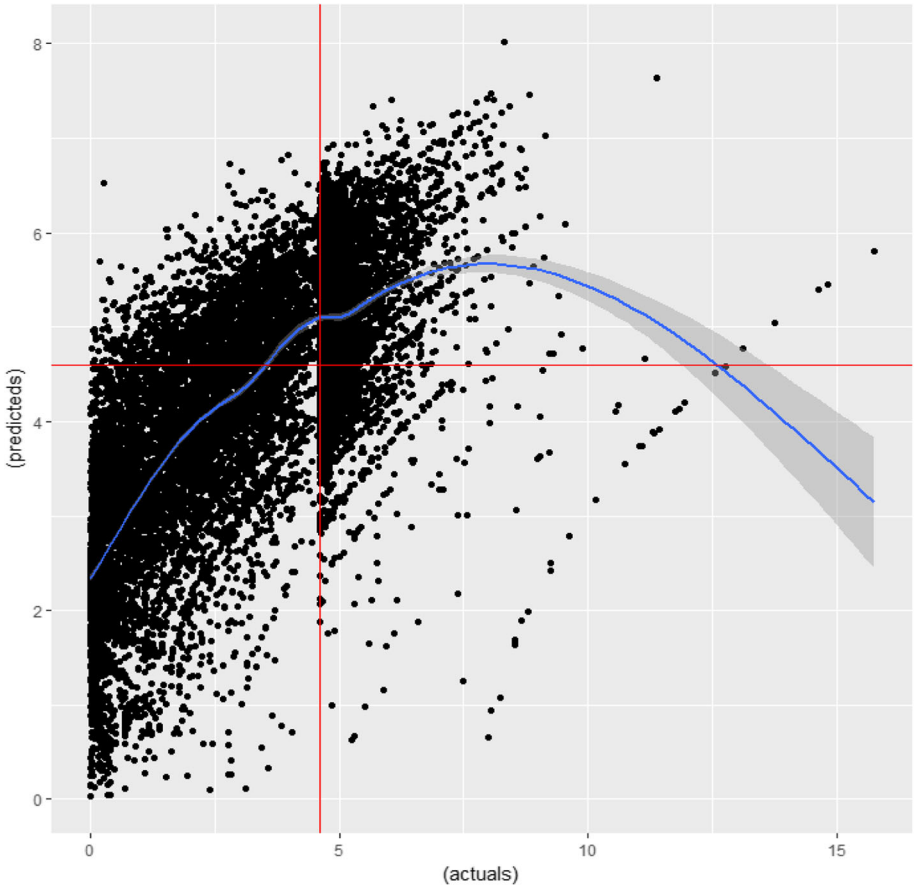

Fig. 11 Predicted scatterplot of predicted values vs actuals separated by 100% of goal lines

zero. The results indicate the geographic location of a Kickstarter plays a role in its success. The effect of the location where a Kickstarter began can be slight to drastic on the performance of Kickstarter campaigns. The posterior variance for the hyperparameters is narrow compared with the overall range, something that is also true with the fixed effects. This narrow variance can show that the model is most likely not overfit and that the data is being explained in a general way so as to help predict future campaigns that are yet to happen.

Pledged USD to a campaign has the strongest pull on the model while the goal is the second strongest variable in the model. There is also important information on how categories are being explained, such as the Twords factor level four having the most positive pull on the model. This seems to make sense though, as Twords was constructed using words that were common among successful campaigns. Category-wise, comics, design, games, and technology had the biggest positive pull on the model whereas crafts, fashion, and journalism, had a negative but smaller impact on the overall model.

Discussion

The proposed dynamical spatio-temporal model for the data did a fair job of giving promising results. The prediction appears to be reasonably strong with around 80%

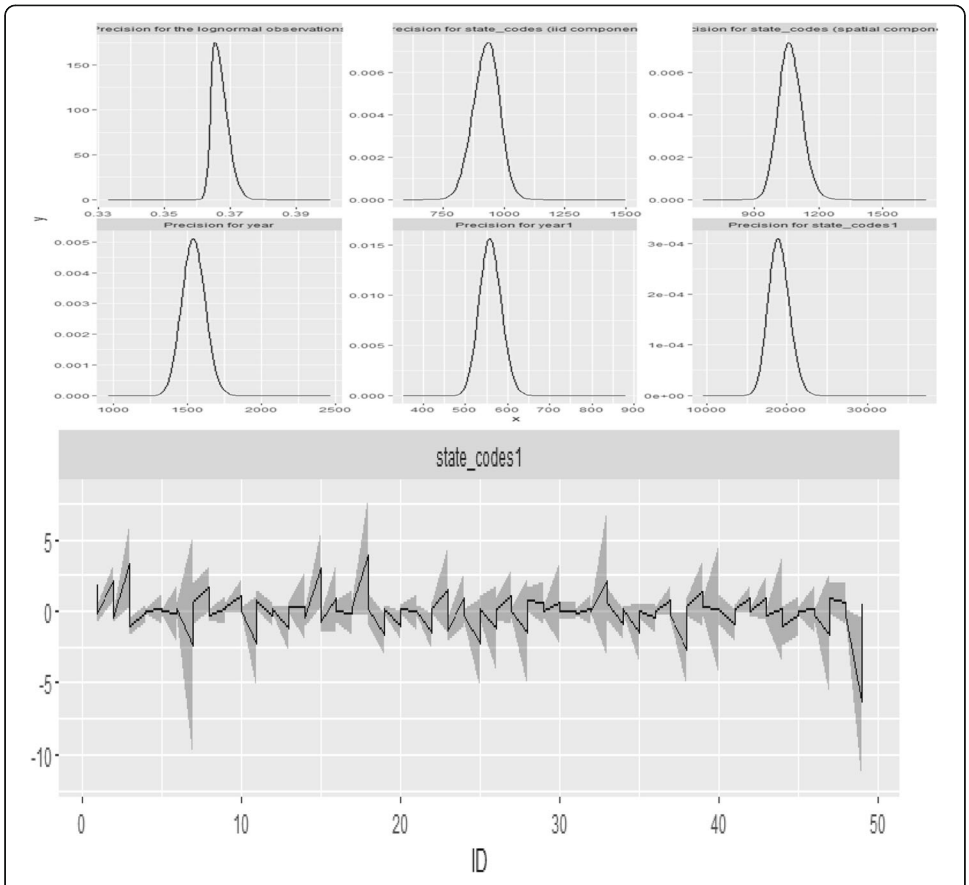

Fig. 12 Top and middle rows show posterior marginal distributions of the model hyperparameters and the bottom row shows the posterior means (black solid line) and 95% pointwise posterior intervals (shaded) for individual states

prediction rates. As far as sheer prediction goes, these models are performing strongly and can give entrepreneurs some insight into how well their Kickstarter campaign may perform based on the variables that they can know beforehand. The presence of type IV interaction showed that a geographic location is not just affected by its neighbors but also by time in its location. This supports the dynamic impact of the geographic location of a Kickstarter on the success of a Kickstarter campaign when it came to prediction.

The random forest algorithm was applied to the data following the same steps for the sake of comparison. The data was split randomly into four subsets of the same size and the hurdle model was fitted. The random forest algorithm was applied to each step, from predicting the back count and the pledged USD to making the final prediction on the reached funding level. The random forest algorithm was able to predict 81% of the actual values and produced very similar results compared with those of the spatio-temporal modeling. The range predicted by the spatio-temporal model was wider than the scope predicted by the random forest within the distribution of the observed data. The random forest algorithm was biased towards more central values, thus the random forest algorithm underestimated the variance of outcomes as well as it cannot predict

more extreme values. Another notable difference was that the random forest algorithm was more likely to predict failed campaigns more accurately while the spatio-temporal model correctly predicted successes more often.

Conclusion

Crowdfunding is a novel method and potentially disruptive innovation for funding a variety of new entrepreneurial ventures, allowing individual founders of for-profit, cultural, or social projects to request funding from many individuals in online communities, often in return for future products or equity. Today, crowdfunding is becoming a major way for entrepreneurs to achieve their dreams. Crowdfunding is mostly viewed from the entrepreneurial perspective as financing including startup capital, one of the most critical of resources required for new ventures to succeed. It is still unknown to scholars and people who use crowdfunding services what makes for a truly successful drive to obtain funding and whether the crowdfunding efforts reinforce or contradict existing theories about the dynamics of successful entrepreneurial financing and the general distribution and use of crowdfunding mechanisms.

This is the first study in the literature using spatio-temporal modeling to understand the dynamics of a successful crowdfunding effort as well as the dynamical impact of geographical locations. Employing spatio-temporal modeling is able to mitigate unobserved spatial confounding when estimating the effect of the factor on a successful entrepreneurial financing, evaluate the impact of the geographic location of a Kickstarter and predict unknown values at unmeasured locations and at future times. Our study involves crowdfunding data collected with explanatory variables at spatial locations from 2009 to the most recent 2017. One distinctive feature of this kind of data is that the data are spatially and temporally indexed to support exploring the hidden dependence structure that is not addressed in the standard machine learning methods through the covariance function of a stochastic process in the spatio-temporal model. The covariance function kernel is essential for the prediction of value at an unobserved location or time. Modeling the covariance function appropriately may improve the efficiency of the estimation of the determinants of a successful crowdfunding campaign and offset the effects of the unobserved sociocultural traits that may affect the determinant under investigation. The spatio-temporal model includes the two components, a systematic component with available explanatory variables and the spatio-temporal correlation component, and how the two components interact to produce reliable forecasts. Such models can thus be reliably used to produce maps and to identify regions (problem or success areas) in the crowdfunding campaign where, for example, the level of performance exceeds the permissible level and thus could be of importance to the success of a new project.

This paper presents new results obtained from investigating the Kickstarter campaign data of over ninety-nine thousand projects totaling about 1 billion USD in pledges from 2009 to the most recent 2017 through spatio-temporal modeling. The funding level is used as the outcome of interest in the modeling to associate with dollar pledged and backer count that reflect the underlying signals of project quality. The spatio-temporal component plays an important role in both the type of projects proposed and the sociocultural traits of successful fundraising related to the underlying quality. Evidence from the results was found to support the impact of the geographic location of a

Kickstarter on its success and the associations between the observed project traits and the success of the entrepreneurial effort in conjunction with the spatio-temporal component. These results offer further insight into the empirical dynamics of the emerging phenomenon of online entrepreneurial financing.

Future work will need to be focused on the hurdle in the model, as figuring out what causes a Kickstarter to fail outright with zero backers seems to be a big obstacle to entrepreneurs and deserves further analysis. It is also necessary to look into the spatial and temporal components and focus on what the fixed effects do through time at each location.

Abbreviations

USD: US dollar; INLA: Integrated Nested Laplace Approximation; DIC: Deviance information criterion; MCMC: Markov chain Monte Carlo

Acknowledgements

The authors would like to thank the reviewers who contributed significantly to the improvement of the article.

Authors' contributions

The study was jointly conceived and designed by Dr. HY and Dr. HH. Dr. HY and Dr. HH revised the article critically for important intellectual content. CW performed the statistical computing under the supervision of Dr. HY. All authors read and approved the final manuscript.

Funding

University of Northern Colorado Fund for Faculty Publications.

Availability of data and materials

The scraped data are available at https://webrobots.io/kickstarter-datasets/.

Competing interests

The authors declare that there are no competing interests.

Author details

[1]Department of Applied Statistics and Research Methods, University of Northern Colorado, Greeley, CO 80639, USA. [2]School of Information, University of South Florida, Tampa, FL 33620, USA.

Received: 7 December 2019 Accepted: 28 May 2020
Published online: 31 July 2020

References

Agrawal, A., Catalini, C., and Goldfarb, A. (2010) The geography of crowdfunding. *SSRN Electronic Journal*.

Banerjee, S. Carlin, B. P. and Gelfand, A. E. (2004) Hierarchical modeling and analysis for spatial data. Boca Raton: Chapman & Hall/CRC.

Belleflamme, P., Lambert, T., & Schwienbacher, A. (2014). Crowdfunding: Tapping the right crowd. *Journal of Business Venturing, 29*(5), 585–609.

Bivand, R., Gómez-Rubio, V., & Rue, H. (2015). Spatial data analysis with R-INLA with some extensions. *Journal of Statistical Software, 63*(20), 1–31.

Blangiardo, M. and Cameletti, M. (2015) Spatial and spatio-temporal Bayesian models with R-INLA. Chichester: John Wiley & Sons, Ltd.

Burtch, G., Ghose, A., and Wattal, S. (2011) An empirical examination of the antecedents and consequences of investment patterns in crowd-funded markets. *SSRN Electronic Journal*.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal, 21*(1), C1–C68.

Cressie, N., and Wikle, C. K. (2011) *Statistics for spatio-temporal data*. John Wiley & Sons.

Ferkingstad, E., Held, L., & Rue, H. (2017). Fast and accurate Bayesian model criticism and conflict diagnostics using R-INLA. *Stat, 6*(1), 331–344.

Frank, M. W. (1998). Schumpeter on entrepreneurs and innovation: A reappraisal. *Journal of the History of Economic Thought, 20*(4), 505–516.

Hu, M. C., Pavlicova, M., & Nunes, E. V. (2011). Zero-inflated and hurdle models of count data with extra zeros: Examples from an HIV-risk reduction intervention trial. *The American Journal of Drug and Alcohol Abuse, 37*(5), 367–375.

Lasinio, G. J., Mastrantonio, G., and Pollice, A. (2013). Discussing the "big n problem". *Statistical Methods and Applications, 22*(1),97–112.

Mitra, T. and Gilbert, E. (2014) The language that gets people to give: phrases that predict success on Kickstarter, proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing 2014, 49-61.

Mollick, E. (2014). The dynamics of crowdfunding: An exploratory study. *Journal of Business Venturing, Volume, 29*(1), 1–16.

Opitz, T. (2017). Latent Gaussian modeling and INLA: A review with focus on space-time applications. *arXiv preprint arXiv:1708.02723*.

Ordanini, A., Miceli, L., Pizzetti, M., & Parasuraman, A. (2011). Crowdfunding: Transforming customers through innovative service platforms. *Journal of Service Management, 22*(4), 443–470.
Rue, H., & Held, L. (2005). *Gaussian Markov random fields: Theory and applications. Monographs on statistics & applied probability*. Boca Raton, FL: Chapman and Hall/CRC.
Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistics Society: Series B (Statistical Methodology), 71*(2), 319–392.
Rue, H., Riebler, A., Sorbye, S. H., Illian, J. B., Simpson, D. P., & Lindgren, F. K. (2017). Bayesian computing with INLA: A review. *Annual Review of Statistics and Its Application, 4*, 395–421.
Schwienbacher, A. and Larralde, B. (2010) Crowdfunding of small entrepreneurial ventures. *SSRN Electronic Journal*.
Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. *Economic Geography, 2*(46), 234–240.
Vuong, Q. H. (2016). Impacts of geographical locations and sociocultural traits on the Vietnamese entrepreneurship. *SpringerPlus, 5*(1), 1189. https://doi.org/10.1186/s40064-016-2850-9.

Publisher's Note