Estimating Contagion Rates in Kickstarter Twitter Cliques

Md Salman Ahmed
Virginia Tech - Department of Computer Science
Blacksburg, VA
ahmedms@vt.edu

Davon Woodard
Virginia Tech - Department of PGG,
Urban Computing Certificate,
and Global Forum on Urban Resilience
Blacksburg, VA
davon@vt.edu

ABSTRACT

In recent years, crowdfunding has emerged as a popular community-based, micro-financing model for entrepreneurs, artists, and activists alike to bring their respective dreams into fruition. Successful campaigns, those which meet their financial goals, bring with them not only the financial utility for the creator, but also social utility for the backers. As such, much attention has been paid in recent years to predicting which campaigns will have the greatest likelihood of success. Rooted in this historical research, this project adapts the model of social media exposure curves of hashtags by Romero, Meeder and Kleinberg [11] to understand contagion spread in Kickstarter crowdfunding campaigns through its twitter network.

CCS CONCEPTS

• **Exposure Curves**; • **Crowdfunding**; • **Social Network Analysis**;

KEYWORDS

Crowdfunding Projects, Information Diffusion, Social Media Coverage, Social Media Exposure, Clique, Twitter Data, Kickstarter

1 INTRODUCTION

Crowdfunding has emerged as a popular crowd-sourced micro-financing mechanism for raising funds to bring entrepreneurial ideas into fruition. In a crowdfunding platform, entrepreneurs (***creators***) submit their ideas or projects along with the project description, goals, prototypes, and rewards. Creators seek people (***backers***) to support the ideas or projects by investing in these new ventures. And, ***promoters***, some times the creators themselves, some times an external person/entity run campaigns to promote the ideas across social networks. Though the reward-based platforms (e.g. Kickstarter, RocketHub, Fundable, and Indiegogo) are the most popular platforms for raising and investing money in new ventures, some other platforms such as the scientific crowdfunding (e.g. experiment.com and crowd.science), donation-based crowdfunding (e.g. GoFundMe and GiveForward), and equity-based crowdfunding (e.g. CrowdCube, EarlyShares, and Seedrs) also raise funds for different purposes. Among these platforms, Kickstarter is the most popular platform due to its "all-or-nothing" funding policy. According to the policy, the creators receive all the pledged amount only if the amount meets or exceeds the goal, otherwise, the respective donated amounts are remitted to the backers.

The micro-financing form of the crowdfunding idea is not only helping new dreams to become reality, but also helping the world's economy by creating new jobs and opportunities. For example, CrowdExpert reported that the crowdfunding platforms raised approximately $2.1 and $3.5 to $4 billion dollars in 2015 and 2016 respectively [6]. World Bank also predicted that the crowdfunding platforms can add as much as $96 billion dollars in the world economy in a single year by 2025 [6]. The potential of securing critical funding through non-traditional funding mechanisms, which often are entrepreneurial roadblocks, motivates creators to submit their ideas through crowdfunding platforms. Many projects such as Finding Vivian Maier film, Pebble E-paper watch, Coolest Cooler, etc. have raised millions of dollars and been turned into reality [3]. However, while a few projects raise millions, most of the projects fail to reach their fundraising goal. For example, only 31 percent project in Kickstarter reached their fundraising goals in 2015 [4]. Failures of projects in crowdfunding platforms draw researchers' attention for developing analytical models for determining the projects' success rate.

While data limitations limited this research from focusing on improving the prediction of success through the inclusion of the stickiness and persistence, this project focuses estimating through the simple and complex contagion method the rate of spread (infection) of a campaign through its network.

2 COMMENTS ADDRESSED

2.1 New Comments

(1) **Inadequate Dataset**. In the very later stage of our project, we realized that we do not have adequate data to develop our models. So we made some assumptions. For example, we needed the Twitter user network for both of our simple and complex models. So, we made an assumption that the user network for our models is a clique.
(2) **Scarcity of associated tweets**. As you expressed the concerns regarding the relevancy of tweets, we actually experienced the scarcity of tweets in terms of both relevancy and quantity. Out of the 18k crowdfunding projects, our crawling program was able to fetch tweets for around 10k projects. However, only 3k successful projects out of 10k projects had enough data to take part in the average exposure curve calculation for the successful projects. Similarly, only 1.5k failed projects took part in the average exposure curve calculation for the failed projects.
(3) **Topic Based Literature Review**. We compressed our literature review a bit to condense it.

2.2 Old Comments

(1) **Vagueness of the term 'temporal data'**. In terms of the temporal data, we tried to express the social media coverage

and social media exposure. We discuss more details how the social media coverage and social media exposure can be quantified in section **??**.

(2) **Unclearness of dataset**. To evaluate our goal, we have collected two sets of dataset. The first dataset includes the descriptive features of projects and the second dataset includes the social media data. We discuss more details of the datasets in the Dataset section.

(3) **Availability of dataset**. A doubt was expressed in terms of the availability of the data to compute the stickiness and the exposure curve. Since we have collected the necessary data and associated the data with the Kickstarter projects, we don't have any issue to compute the stickiness and exposure curve. However, we slightly modified our variables. Instead of computing stickiness, we now compute the coverage. We discuss details of these computations in Exposure Measurement section.

(4) **Topic Based Literature Review**. We added topics to our literature review. However, the content of each topic was not condensed in this version. Surely, we will condense the literature review section in our final report.

3 RELATED WORK

The expanding interest in this area arises from both the unique structural challenges inherent in crowdfunding data, as well as the significance of the outcomes for both the creators and backers. Predicting the success of crowdfunding campaigns has become an increasingly popular research focus in both the computer and social sciences.

3.1 Foundational Research

In creating a broad understanding of campaign prediction, Cordova, et. al, [2] provided an interdisciplinary perspective of the determinants of success for crowd-sourced funding project success, focusing specifically on early stage technology start-ups from four crowd-funding platforms Kickstarter, Ulule, Eppela, and Indiegogo.

In their work they analyze 1,127 cases across seven variables: log crowdfunding goal, number of funders, log mean contribution amount, project duration, log mean of daily contributed amount, location, number of project updates, number of donor comments, and type of funding, characterized as all or nothing. A simple probit regression of the data subsets found that project request amount, contribution frequency and campaign length were important factors in predicting campaign success [2].

The work of Cordova was complemented by researchers, Lu, et. al, who provided a broad understanding of the intersection of social media and crowdfunding as they progressed through their research by observing a few interesting phenomena [8]. The researchers found the number of promoters in social media does not predict the number of backers, rather the number of backers depends on the quality of the promoter activities/campaigns. Their second observation is that the popularity of a project among backers fades away until the approaching of the fundraising deadline of the project. So, quality campaigns in right times can keep the project's popularity throughout the fundraising duration, which in turns increases the likelihood of the project to be successful. Keeping these two observations in mind, the authors extracted features from Kickstarter dataset and the corresponding creators', backers', and promoters' social media accounts to train the dataset using machine learning techniques (e.g. singular value decomposition and logistic regression) to predict the project's success and the amount of fund the project is going to receive.

3.2 Feature-Based Predictions

As some research focused on techniques, other research focused on feature selection to create better models of prediction. Research in the field began to explore various predictive tools to increase the predictive power of social media data. In one study, researchers in [1] extracted a unique feature called SMOG grade [9] from the project description and the project related tweets. Then they developed and compared their predictive models using Naive Bayes, Random Forest, and AdaboostM1 machine learning algorithms. They found that their predictive models can achieve up to 76.4 percent accuracy. The authors also analyzed the lifetime and profile of creators and backers on Kickstarter and found three interesting facts: 1) experienced creators are more successful than the first time creator; 2) backers tend to fund the project with shorter duration than longer; and, 3) the creators who become successful have large number of Facebook friends.

3.3 Social Feature-Based Predictions

In "Launch Hard or Go Hard," Etter, Grossglauser and Thiran [5] propose three models using both "direct and social features" of Kickstarter campaigns to predict the success of campaigns, as measured by the realization of their financial goals, at their termination of their fundraising period. The first model, using a time series step of pledged money, the second using secondary information gathered from social media, particularly tweets, and, third, a combined approach.

In parallel, they also scraped Kickstarter for new campaigns with associated twitter data and trained their financial model using KNN (k=25) and Markov chain (N_{m30}). For the social media model, they evaluated: (1) the number of tweets and retweets; and, (2) a some backer-based estimations. Using an SVM on both data sets, they found a higher than base (66 percent) predictive capability early on in the time series, however, the capability dropped significantly towards the end of campaign life [5].

Here, the combined model, balanced the shortcomings of the respective models and provided higher end of stage success predictability. Most importantly, within the first four hours (4 percent) of launch, the combined model was able to predict campaign success with an accuracy rate higher than 76 percent [5].

4 DATASET

To estimate the rate of spread (infection) of a campaign through its network using the simple and complex contagion method, we collected two set of datasets (e.g. Kickstarter dataset and Twitter dataset). The first set of data include the regular project features and the second set of data deal with the social exposure of the projects.

Table 1: Summary of datasets

Criterion	Before cleaning	After cleaning
Total projects	18,142	9,812
Unique project creators	16,827	8573
Number of projects having tweets	10,249	9,812
Total collected tweets for projects	162,375	162,374
Unique users tweet for projects	98,200	98,200

4.1 Kickstarter Dataset

To extract the regular crowdfunding project features, we utilize the dataset from Kickstarter. We collected a dataset of projects that were submitted to Kickstarter within the duration from December 2013 to June 2014. The dataset is a comma separated file that has more than 18k rows. Each row of the dataset presents the features of a project. The features include, but not limited to the creator's name, project's URL, project's category, duration, total goal amount, total pledged amount, amount of reward-points, number of backers, number of Facebook shares, number of times the project gets updated, does the project have video demonstration, etc.

4.2 Twitter Dataset

To collect the Twitter data for our projects, we utilized a Twitter data crawling program [10] from the Internet and modified the program to meet our needs. We ran the program for three days to crawl Twitter data for each of our projects. To get the Twitter data for a project, we fed the keywords as a search query by appending the word 'kickstarter' with the project name. The 'kickstarter' word ensures that all the tweets that we are collecting are somehow related to the Kickstarter projects. In case of a small project name, i.e. if the project name consists of only two words or less (e.g. Cold Again, The End, Scout, Father, etc.), we appended the last part of the project URL with the keywords. The Twitter data crawling program is designed to retrieve the associated tweets, dates of the tweets, names of the twitters, and the tweet links. We manually verified the relevancy of tweets that are associated with a project by randomly sampling some tweets. We also manually compared the total number of tweets that the crawling program gave us vs the number we get by searching on the Twitter website. The difference between the two number is very low (e.g. ±1 or ±2).

4.3 Data Cleaning

Out of the 18k projects, the crawling program was able to fetch more than 162k rows of Twitter data for around 10k projects where each row includes the name of the user who tweeted, date of a tweet, how many times a tweet gets retweeted, and the link of a tweet. A few dates could not be parsed due to invalid date format error. So, we cleaned the data by filtering out the corresponding rows of these dates. Table 1 illustrates the summary of the raw and cleaned data.

Table 2: Summary statistics of the crowdfunding projects. Raw and cleaned feature values are given for overall, successful and failed projects

Features	Overall		Successful projects		Failed Projects	
Avg. number of updates	3	4	5	6	2	2
Avg. number of comments	32	46	64	75	3	4
Avg. goal amount	27330	32086	9620	10710	45065	64065
Avg. duration	31	31	29	30	32	33
Avg. number of Facebook shares	392	514	670	735	132	203
Projects having video demo	83%	88%	89%	91%	77%	83%
Avg length of description (# words)	656	726	717	757	604	690
Avg. length of risks & challenges (# words)	128	135	130	133	127	138

4.4 Data Set Findings

To illustrate the impact of different features on a project's success, we analyze and derive the summary statistics for the Kickstarter dataset. Table 2 illustrates the summary statistics for the cleaned Kickstarter dataset. Since the original Kickstarter dataset had around 18k projects, we compare the statistics of the cleaned dataset (that contains around 10k projects) with the original dataset to identify how much and of what extent data we lost after cleaning the data. Out of the 18k projects, we found that the project success rate is around 48%. Most successful projects were submitted to Film & Video and Music category. Also, the most successful projects were submitted from Los Angeles, New York, and London. From the summary statistics, we found that a few features such as comments, updates, and Facebook shares have a good correlation with the successful projects. Since these features have a good correlation with the successful projects, we hope that the social media exposures may have a good predictive power for predicting successful projects.

5 PROPOSED MODELS

This project attempted to adapt the work of Romero, et. al.[11]. In this work the team sought to understand how the stickiness and persistence of the tweet diffusion through a Twitter user network could impact the success of crowdfunding projects. However, due to the limitations of our twitter dataset (e.g., full Twitter user network, complete set of nodes and neighbors were not known), we sought to understand the infection rate in as detailed in two models, a simple SI contagion model and a complex contagion model.

5.1 Simple Model

The simple model used for this project adapts the original epidemic SIR model to our data set. In the SIR model it is assume that $N = S + I + R$ [7]. Here the total population is the summation of those who are susceptible, infected and those have been infected and recover. The rates of change in each in calculated as:

$$\frac{dS}{dt} = -\beta SI$$
$$\frac{dI}{dt} = \beta SI - \gamma I$$
$$\frac{dR}{dt} = \gamma I$$

In this formulation, β represents the contact rate for infection and βSI is the rate of infection in respective populations. For our research, there have been a few necessary assumptions made in order apply this framework to our data set. First, due the lack of information on complete network structure, we assume that each project is a closed clique, where $N = 2 * tweeters_{unique}$. Second, we assume no recovery, and that is once a user tweets there is a permanent infection. In this way, we eliminate the recovery rate, γI, as well as the rate of loss of infection, γI. Given these assumptions, we attempt to estimate $\frac{dI}{dt} = \beta SI$.

5.2 Complex Model

To build a generalized complex contagion model, we determine the average exposure curves for all successful and failed projects. We follow the techniques described by Romero et. al. [11] to determine the exposure curve for each project. Since we do not have the exact Twitter user network, we assume the network as a clique and total nodes of the clique is twice the total number of unique Twitter users who tweet or retweet for the project. For example, if twenty unique Twitter users tweet for a project within a duration of forty days, then we consider the user network as a clique containing the twenty users (who tweet) as well as another twenty users (who doesn't tweet). We discuss the assumptions in details in the discussion section to describe why and how we make the assumptions.

We consider an interval of one day to determine the newly infected (infected in the sense that they tweeted) users who are exposed to k users. Here, k ($k = 1, 2, 3, 4, 5...$) indicates the number of Twitter users who have already tweeted for a project in the beginning of a time step t_1. In the beginning of t_1 time step, we also calculate the total number of k-exposed users (E_k). Then we determine the total number of unique Twitter users (I_k) from E_k who tweet for the project within a duration of one day. We calculate the probability $P(k)$ that a user will tweet after being k-exposed by dividing the I_k by E_k. We calculate the $P(k)$ curves for all the successful and failed projects. Then we determine the average $P(k)$ curves for all the successful and failed projects. Figure 1 illustrates the exposure curves.

From these two exposure curves, we can see that the probability of infection is being declined exponentially. Since each of the exposure curves is being declined exponentially, we can fit the curve to an exponential decay function to get the decay rate. We fit equation 1 to the curves. The green dots in Figure 1 (both a & b) represent the approximated (fitted) curve. We calculate the decay rates for both successful and failed projects. The decay rates for the successful and failed projects are 0.11360768 and 0.12637707, respectively.

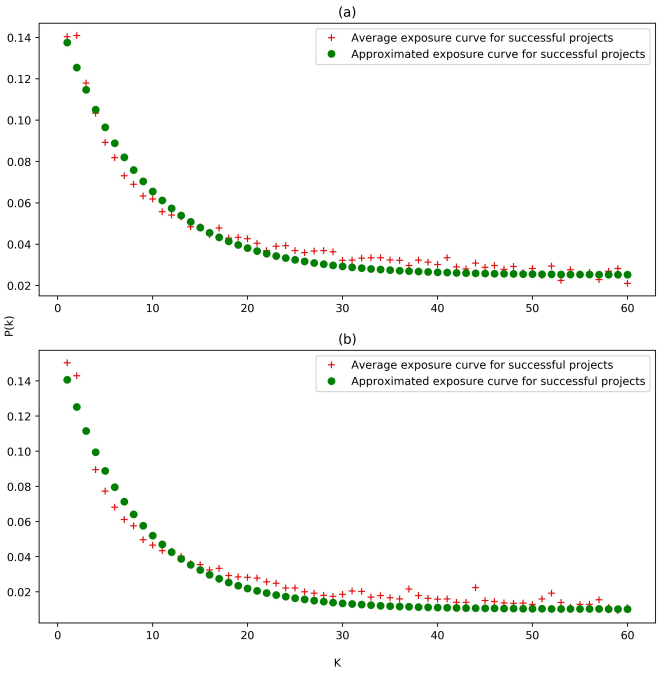

Figure 1: Average exposure curves: (a) for successful projects and (b) failed projects

$$y = \delta e^{-\lambda x} + c \tag{1}$$

Here δ represents probability of initial infection, λ represents the decay rate, and c is a constant.

Since we consider the user network in a clique setting, a differential equation can represent the infection rate. For simple SI model, the following equation 2 represents the infection rate.

$$\frac{dI(t)}{dt} = \beta S(t) I(t) \tag{2}$$

Where β is the probability of infecting susceptible nodes, $S(t)$ is susceptible nodes at time t, and $I(t)$ is the infected nodes at time t. Since we can determine the decay rate from the generalized exposure curve, a generalized complex contagion model can be modeled to represent the infection rate by replacing the probability β by $P(k)$ where $P(k) = \delta e^{-\lambda k} + c$. The following equation represents the infection rate for the generalized complex contagion model. The differential equation can be further modified to replace the value k by $I(t-1)$

$$\begin{aligned}\frac{dI(t)}{dt} &= P(k) S(t) I(t) \\ &= (\delta e^{-\lambda k} + c) S(t) I(t) \\ &= (\delta e^{-\lambda I(t-1)} + c) S(t) I(t)\end{aligned}$$

Where c is a constant and λ is the decay rate, 0.11360768 and 0.12637707, for successful and failed projects respectively in our case.

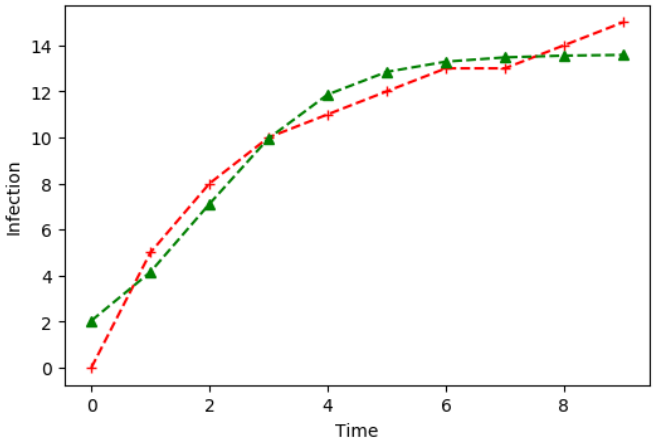

Figure 2: Fitting the infection curve. Red marked line represents the actual infection and the green curve represents the fitted curve. Time represents day

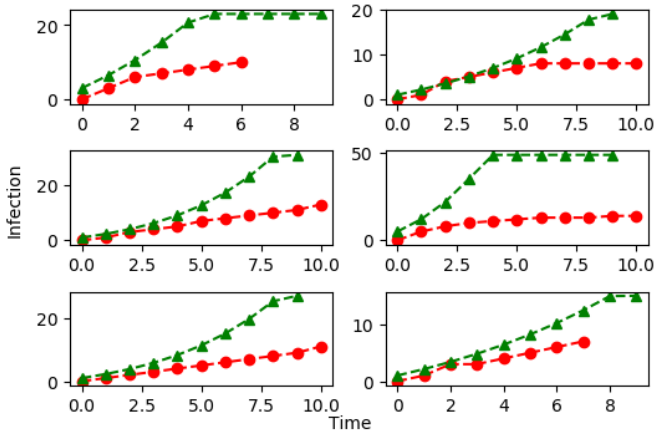

Figure 3: Actual (red) versus predicted (green) infection for six successful projects using the simple model. Time represents day

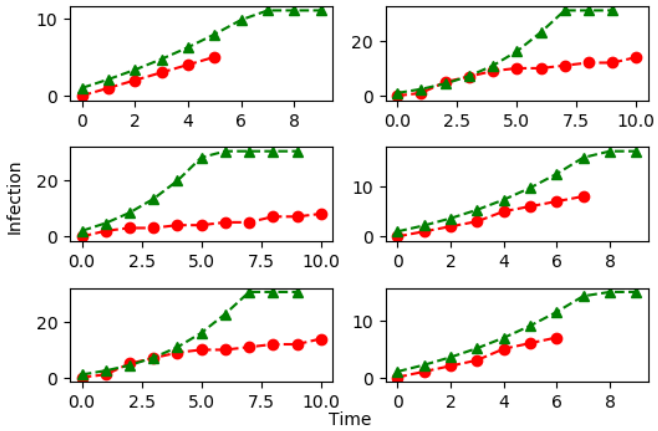

Figure 4: Actual (red) versus predicted (green) infection for six failed projects using the simple model. Time represents day

6 EVALUATION AND RESULTS

6.1 Simple Model Evaluation

We undertook the following evaluation technique to evaluate the simple model, i.e. to approximate the β.

In this technique, we first integrate the infection rate differential equation, $\frac{dI(t)}{di} = \beta S(t)I(t)$, with respect to t within an interval of $[0, t]$ to get the infection function, $I(t)$ (equation 3) [12].

$$I(t) = \frac{NI_0e^{\beta Nt}}{N + I_0[e^{\beta Nt} - 1]} \tag{3}$$

Where N is the total population, β is probability of infection, and I_0 is the initial infection.

We calculate the actual infection curve for all successful and failed projects and fit equation 3 to approximate beta values. Figure 2 illustrates the fitted curve. We average the beta values for all successful and failed projects to get an average beta value for successful projects and another for failed projects. Once we have the average beta values (in our case 0.00924 for successful projects and 0.01198 for failed projects), we compare the actual infection versus the approximated infection over 10 days. Figure 3 illustrates the actual vs. approximated infection for six randomly selected successful projects, and Figure 4 for six randomly selected failed projects.

Using this evaluation technique, we see that the actual and predicted infections are almost similar up to 5 days. After 5 days, we see a big difference between the actual and approximated infections. We discuss the possible reasons for this approximation error in our discussion.

6.2 Complex Model Evaluation

To validate the complex contagion model, we predict the infection of all the successful and failed projects for the first 10 days from the project starting date using the complex contagion model. Figure 5 & 6 illustrate the actual and predicted infection for six randomly selected successful and six randomly selected failed projects, respectively. However, we see a big difference between the predicted and actual infections over time. One obvious reason for this difference is the inaccurate Twitter user network. Another reason is the scarcity of twitter data. Out of the 18k crowdfunding projects, our crawling program was able to fetch tweets for around 10k projects. However, only 3k successful projects out of 10k projects had enough data to take part in the average exposure curve calculation for the successful projects. Similarly, only 1.5k failed projects took part in the average exposure curve calculation for the failed projects.

7 CONCLUSION AND DISCUSSION

Through the simple contagion method and complex contagion methods, this project sought to understand the spread of contagion through Twitter cliques associated with Kickstarter fundraising

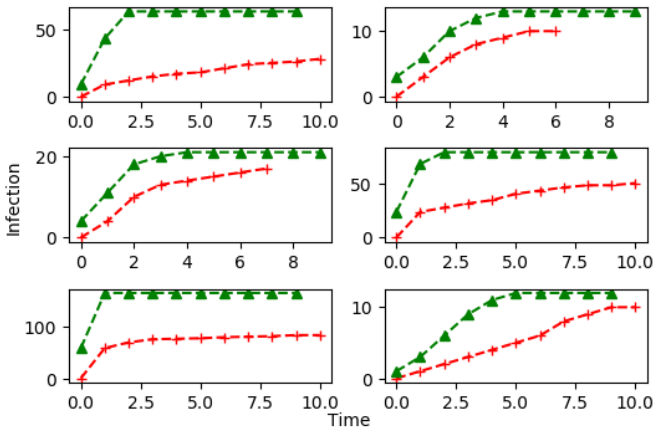

Figure 5: Actual (red) versus predicted (green) infection for six successful projects using the complex model. Time represents day

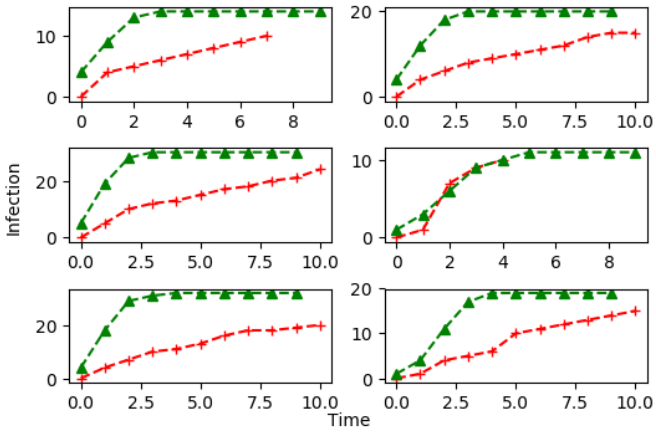

Figure 6: Actual (red) versus predicted (green) infection for six failed projects using the complex model. Time represents day

campaigns. Using the simple contagion model, we were able to approximate the Twitter data to some extent. As noted, when this model is run without respect to time, and solely at the evolution of the contagion through the clique network as a function of the nodes, the bell-shaped curve is approximated denoting the increasing at a decreasing rate of infection of new nodes, as the total population becomes infected.

Also, as we noted before, we see a big difference between actual and predicted infection using the generalized complex contagion model. The primary reason for that is the inaccurate Twitter user network. Our Twitter data crawling program was designed to retrieve only the tweets and usernames associated with Kickstarter fundraising campaigns. However, in the later stage of our work, we needed the actual Twitter user network. Since we didn't have enough data to build a Twitter user network, we assumed the user network as a clique of the unique Twitter users who tweeted for a project. This assumption is not true for real network. The second reason for the approximation error is the inaccurate population. Due to the scarcity of data, we assumed that the user network is a clique of the unique Twitter users who tweeted for a project. Since the clique contains only the users who tweeted, the calculation of our exposure curve was not right. So, we doubled population size, i.e., we assumed a clique of size twice the unique Twitter users. Though we were able to calculate an average exposure curve using this clique size, in reality, the network may not be a clique and the network size could be much larger.

Finally, the approach we have undertaken in this research is beneficial for the backers' community to identify the crowdfunding projects that they could invest on. So, the possible future research direction could be the refinement of the simple and complex model using more social data. The simple and complex generalized models also need some sort of threshold analysis to accurately predict the future activity of crowdfunding projects.

REFERENCES

[1] Jinwook Chung and Kyumin Lee. 2015. A long-term study of a crowdfunding platform: Predicting project success and fundraising amount. In *Proceedings of the 26th ACM Conference on Hypertext & Social Media*. ACM, 211–220.

[2] Alessandro Cordova, Johanna Dolci, and Gianfranco Gianfrate. 2015. The Determinants of Crowdfunding Success: Evidence from Technology Projects. *Procedia - Social and Behavioral Sciences* 181 (2015), 115–124. https://doi.org/10.1016/j.sbspro.2015.04.872

[3] Entrepreneur. 2017. The 10 Most Funded Kickstarter Campaigns Ever. (2017). https://www.entrepreneur.com/article/235313 Retrieved: 2017-10-18.

[4] Entrepreneur. 2017. Less Than a Third of Crowdfunding Campaigns Reach Their Goals. (2017). https://www.entrepreneur.com/article/269663 Retrieved: 2017-10-18.

[5] Vincent Etter, Matthias Grossglauser, and Patrick Thiran. 2013. Launch hard or go home!: Predicting the Success of Kickstarter Campaigns Vincent. *Proceedings of the first ACM conference on Online social networks - COSN '13* (2013), 177–182. https://doi.org/10.1145/2512938.2512957

[6] Forbes. 2017. The Rise Of Investment Crowdfunding. (2017). https://www.forbes.com/sites/adigaskell/2016/03/15/the-rise-of-investment-crowdfunding/#12e87f1a4d9b Retrieved: 2017-10-18.

[7] Jordan Hasler. 2013. SI Epidemics Model. (2013).

[8] Chun-Ta Lu, Sihong Xie, Xiangnan Kong, and Philip S Yu. 2014. Inferring the impacts of social media on crowdfunding. In *Proceedings of the 7th ACM international conference on Web search and data mining*. ACM, 573–582.

[9] G Harry Mc Laughlin. 1969. SMOG grading-a new readability formula. *Journal of reading* 12, 8 (1969), 639–646.

[10] Get Old Tweets Programatically. 2017. (2017). https://github.com/Jefferson-Henrique/GetOldTweets-python Retrieved: 2017-11-12.

[11] Daniel M. Romero, Brendan Meeder, and Jon Kleinberg. 2011. Differences in the mechanics of information diffusion across topics. *Proceedings of the 20th international conference on World wide web - WWW '11* (2011), 695. https://doi.org/10.1145/1963405.1963503

[12] SIS Solution. 2017. (2017). http://mysite.science.uottawa.ca/rsmith43/MAT4996/Epidemic.pdf Retrieved: 2017-12-10.