Collection, Exploration and Analysis of Crowdfunding Social Networks

Miao Cheng Anand Sriramulu Sudarshan Muralidhar
Boon Thau Loo Laura Huang Po-Ling Loh
University of Pennsylvania
{miaoch,anandsri,smural,boonloo}@seas.upenn.edu,
{huangla,loh}@wharton.upenn.edu

ABSTRACT

Crowdfunding is a recent financing phenomenon that is gaining wide popularity as a means for startups to raise seed funding for their companies. This paper presents our initial results at understanding this phenomenon using an exploratory data driven approach. We have developed a big data platform for collecting and managing data from multiple sources, including company profiles (CrunchBase and AngelList) and social networks (Facebook and Twitter). We describe our data collection process that allows us to gather data from diverse sources at high throughput. Using Spark as our analysis tool, we study the impact of social engagement on startup fund raising success. We further define novel metrics that allow us to quantify the behavior of investors to follow and make investment decisions as communities rather than individuals. Finally, we explore visualization techniques that allow us to visualize communities of investors that make decisions in a close-knit fashion vs looser communities where investors largely make independent decisions. We conclude with a discussion on our ongoing research on causality analysis and new community detection algorithms.

1. INTRODUCTION

Technology startups are an important engine for economic growth. Recent trends show that while existing firms contributed to a loss of over 1 million jobs annually, new firms in their first year of existence added over 3 million jobs annually [16]. Examining the characteristics of high-tech startups that contribute to such massive progress is therefore important. Because startups and early-stage organizations are not merely smaller versions of larger companies, prior research has not been able to capture the evolution of entrepreneurial startups using data-driven techniques, and identify the factors that are most predictive of critical outcomes.

Four critical factors contribute to survival and success in the entrepreneurial process: (a) the ability to test and validate product ideas that resonate with customers ("product-market fit"); (b) the ability to raise capital at various stages of growth while the startup searches for the right business model; (c) the ability to advertise and sell products to early adopters as startups seek growth; and (d) the ability to engage constructively with other players within the entrepreneurial ecosystem (e.g., partners, suppliers, and other resource providers). These factors are evolving in light of the increasing use of social media in various aspects of startup operations, ranging from idea generation and validation (crowdsourcing), to startup fundraising (crowdfunding) and marketing (social media marketing). No longer do startups rely solely on traditional sources of funding, such as venture capital and government-sourced funding, or conventional tactics for advertising, marketing, and communicating with outside constituents. Internet technology allows startups to couple traditional approaches with more decentralized and customer-driven approaches.

A critical factor of a startup's success is its ability to disseminate information and market products through social media and other platforms (e.g., via company profiles, social media engagements, and public evidence of funding and public endorsements) may help attract customers and high-quality investors, translating into subsequent product sales and strong financial outcomes. Such activity is not only economically significant, but also presents a rare opportunity for data science researchers. A big data study that analyzes the large amount of publicly available data could lead to informative new predictions and a better understanding of the factors leading to startup success.

In this paper, through the actual collection and analyzing of social media data, we seek to understand the factors that influence one particularly important activity of a startup, namely *fund-raising*. We focus on an emerging financing mechanism that has enjoyed wide popularity called *crowdfunding*. In crowdfunding, a startup uses a portal such as AngelList [1], Fundable [5], or EquityNet [4] to launch a fundraising campaign. Typical investors may pledge small amounts of funding (as little as $1000) for equity. Crowdfunding companies then leverage social media to raise awareness among potential backers. Hence, they release a massive amount of online material concerning opinions on their industry and the background of their team. "Buyers" correspond to accredited investors, who possess the ability to make small risky investments in growing companies. Investors in crowdfunded companies often perform less due diligence (compared to traditional investors), given the small amounts of capital invested and their general lack of expertise.

In particular, this paper makes the following key contributions:

- **Extensible Exploratory platform for data collection and analytics.** We propose an extensible exploratory platform for collecting and managing data from multiple sources, including company profiles (CrunchBase, AngelList) and social networks (Facebook and Twitter). Our big data platform is designed to be highly extensible (e.g., for customizing new data sources and analytics), use scalable backends such as the Hadoop File System (hadoop.apache.org) and Spark

ExploreDB'16, June 26-July 01 2016, San Francisco, CA, USA
© 2016 ACM. ISBN 978-1-4503-4312-1/16/06...$15.00
DOI: http://dx.doi.org/10.1145/2948674.2948679

(spark.apache.org), and in future, will expose familiar user interfaces for social scientists.

- **Large-scale data collection.** We have carried out a systematic data collection process, at collecting various social media data related to startup fundraising. Using public APIs, we have gathered data from AngelList, Twitter (twitter.com), Facebook (facebook.com), and CrunchBase (crunchbase.com). AngelList is a U.S. website for startups, angel investors, and job seekers for startups. AngelList also allows investors to invite other accredited investors to form syndicates for investment. CrunchBase contains information on startups, founders, and their latest rounds of financing. Note that these websites are not only connected to startups and investors via social media, but are also interconnected to each other.

- **Spark-based exploration and analytics.** The collected data are stored in Hadoop File System (HDFS) [8]. We further use Spark [12] for integrating, cleaning, and analyzing our datasets. We report a number of interesting observations related to the effectiveness of social media engagement as a factor in influencing fund-raising success. Companies with a social media presence are 30X more likely to succeed in fundraising, and the percentages are further increased through the active engagements on social media (e.g. frequent tweets and posts, and use of demo videos). There is also suggestion of "herd mentality", where the use of standard community detection algorithms reveal that many investors frequently coinvest in similar companies. We conclude with a discussion of our ongoing research at a longitudinal data capture and study for an in-depth causality analysis, exploratory data analysis based on social psychology hypothesis, and techniques for performing predictions on individual companies based on attributes exhibited during early stages of development.

2. BACKGROUND

We first provide a background introduction to crowdfunding and the data sources that we have used. In the context of startups, crowdfunding is the practice of funding a venture by raising monetary contributions from a large number of people, today often performed via Internet websites. The crowdfunding website that we focus on in this paper is AngelList, given that it is widely used, and that there is an available public API provided for us to collect data. According to its online description [2], AngelList is a US website for startups, angel investors, and job-seekers looking to work at startups. It started as an online introduction board for tech startups that needed seed funding. Over time, the site evolved into one that allows startups to raise funding from angel investors who are accredited. AngelList is now one of the most popular crowdfunding websites in the world.

AngelList allows anyone to register and log in as an independent user. In AngelList, one can serve the role of startup founders, investors, or employees. The website allows companies to publicize its own profiles, launch fundraising campaigns, advertise jobs, and provide links to its social media websites (Twitter, Facebook).

Figure 1a shows the profile page for a startup named *Planetary Resources* in AngelList. A startup's profile page contains many features, including its overview, activity, followers, founders, team, funding and so on. This profile page includes several relevant links, such as the homepages of all the involved people (founders, investors, and employees), the startup's official website, and its Twitter, Facebook, LinkedIn accounts. In this way, AngelList is similar to social media websites, such as Twitter and Linkedin, forming a huge social networking graph in the startups' world.

In addition to AngelList, another relevant source of fund-raising information that we draw upon is CrunchBase [3]. The website includes funds raised by every startup, and provides a public API for retrieving the information. CrunchBase and AngelList automatically synchronizes with each other. Hence, any public information provided for a startup on AngelList can be automatically exported over to its corresponding CrunchBase company profile, and CrunchBase is linking back to AngelList for every company that has an entry in both places.

Figure 1b shows an example profile page obtained from CrunchBase. There are many same features in CrunchBase as in AngelList, such as the basic overview information, founders, investors, followers, timeline activity and so on. CrunchBase also provides the funding details including how much money a startup has raised in each round with exact date and the number of investors. Undoubtedly, this is very critical in measuring the success of a startup ability to fundraise by advertising itself on AngelList. Figure 1b further shows a networking graph of startups, investors, and teams as well as insights like common investors in different startups, and teams' working experience. This information is helpful for us to analyze the reasons behind the success of a startup.

3. DATA COLLECTION AND ANALYTICS

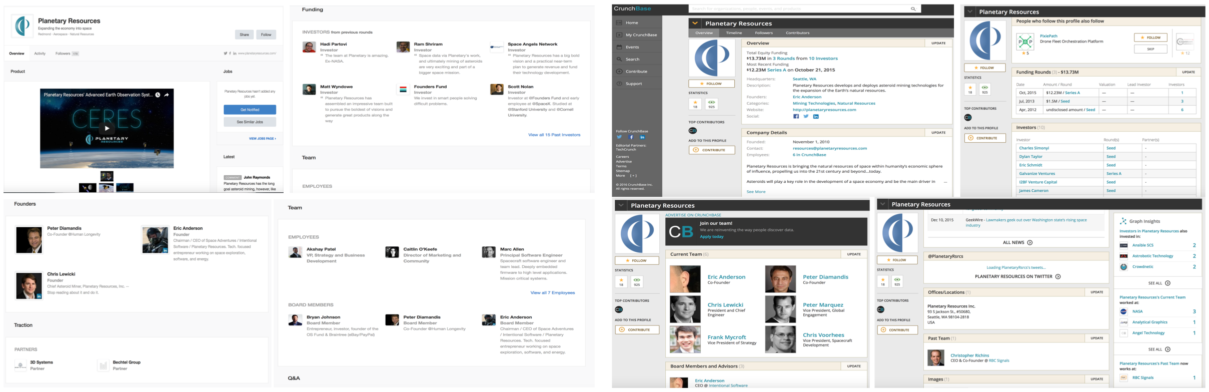
Figure 2 shows our data collection and analytics architecture. A number of high-performance parallel crawlers are used to gather social media inputs from Facebook, Twitter, CrunchBase, and AngelList. We adhere to the Web APIs supplied by each company, and avoid direct web-scraping techniques that may violate company policies. We also provide mechanisms to crawl these sources periodically and track them over time. Though for the purpose of this paper, we focus our initial analysis on a one-time collection of data, and leave the longitudinal study to future work.

The crawled data will be stored in the HDFS. We formulate different social, behavioral, and economic theories. Parallel statistical and machine learning queries can also be directly programmed in Spark to analyze the crawled data. Our platform also allows for external plug-ins, for example, the use of external community detection libraries, which we will describe in Section 5. In future, we plan to provide familiar interfaces to social scientists, so that they can directly validate theories using computational platforms such as R, Matlab, and SPSS. A translation layer will map the theories to Spark queries for execution.

Our crawled data are stored in HDFS as files in the JSON (JavaScript Object Notation) format. JSON is a well-known industry standard for storing documents in a platform-independent manner for analysis. We next run the Spark software over the crawled data. We use Spark primarily for cleaning, extracting and summarizing data from all our social media sources. The processed data is then further processed by Spark's statistical analysis modules, or used as input to a third party tool for further processing.

In the rest of this section, we will describe our data collection process in greater detail, divided into data sources.

AngelList. Our crawl starts from AngelList as our authoritative source of startups. We use the AngelList API to query and download startup information. AngelList's API currently only provides a list of all startups that are currently raising money (about 4000 of them). In order to collect more information about all startups in AngelList's database (including those that have previously raise, or are not yet raising funding), we use an algorithm similar to breadth-first search over graphs. We first collect information on all currently raising startups. We call this set the *frontier*. We next collect a list



(a) **Example AngelList profile page.**

(b) **Example CrunchBase profile page.**

Figure 1: **Screenshots of profile pages.**

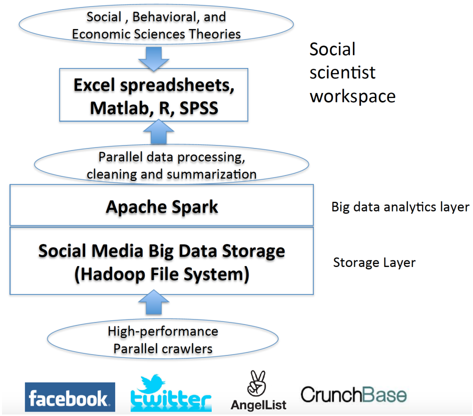


Figure 2: **Extensible Exploratory Platform.**

of all users that are following a startup in the frontier. This set of users becomes the new frontier, and we collect, the set of users followed by all users in the frontier, as well as all startups and users followed by a user in the frontier. As before, we make this newly collected set the frontier, ignoring any startups or users that have been in the frontier before. We repeat this process, increasing our knowledge of the entire AngelList graph in every iteration. After several rounds, we are able to collect more than 700K startups with varying properties.

CrunchBase. AngelList data is incomplete. Specifically, many startups have funding events that, for varying reasons, are not reflected in their AngelList profiles. In order to have as accurate information as possible, we augment our AngelList data with crawled data from CrunchBase. Because CrunchBase funding data does not change very frequently (especially when compared to AngelList), we perform a one-time augmentation to the AngelList data - upon finishing our initial breadth-first search crawl over AngelList, we query CrunchBase for each of the AngelList startups. If the AngelList entry provides a CrunchBase URL, we use the associated CrunchBase entry; if not, we use the CrunchBase search API to find startups with matching names. If the CrunchBase search returns a unique result, we associate that result with the AngelList startup. This allows us to augment a large number of AngelList startups with CrunchBase data.

Facebook. The AngelList dataset includes links to startups' available Facebook and Twitter URLs. We use Facebook's Graph API [7] to extract each startup's profile info out of Facebook Platform. The Graph API provides a low-level HTTP-based API such that when given a certain startup's Facebook URL, we can get its basic profile fields which includes its location, the number of likes, and the recent posts. When calling the Graph API, our Python-based crawler logs into the Facebook as a user, and get a valid access token before querying any data. The access token is at first short-lived, but we've used it to generate a long-lived one through certain procedures including creating a Facebook App. Therefore, our Facebook crawler can work without any limitations.

Twitter. Finally, we use the tweepy python library [13] to call the Twitter RESTful API methods to extract data of the Twitter platform. Because we lack the startup's Twitter account id, we use its URL (gathered from AngelList) to obtain profile info. We extract the startup's Twitter username from its Twitter URL (the string after the last "/" symbol), then use the username as the key to query the data. We extract all the critical information provided by the Twitter RESTful API. The features we get from a startup's Twitter profile include: its created time, its followers count, friends (following) count, listed count, statuses (tweets) count, and its latest status (tweet) with timestamp. Twitter API's rate limit is 180 calls every 15 minutes, and we are also required to use access tokens to reach the data. The tokens are generated by registering Twitter apps, and each twitter user is allowed to register at most five apps in a certain period of time. Hence, we distribute the Twitter crawling job to several machines, using different access tokens, which tackles the rate limit issue effectively.

Using the above mechanism, we downloaded company information from 744,036 companies on AngelList and 10,156 CrunchBase profiles to augment our fund-raising information. Of these AngelList companies, we also collected 37,761 and 70,563 company profiles correspondingly from Facebook and Twitter, based on whether these companies have valid links from AngelList. Also included in the AngelList data is information on 1,109,441 users, of which 47,345 (4.3%), 203,023 (18.3%), and 489,836 (44.2%) identified themselves as investors, founders, and prospective employees respectively.

Figure 3 shows the CDF of investments made by investors. The CDF clearly shows the presence of a long-tailed distribution, where a small number of investors make a large number of investments, Our data reviewed that on average, each investor *follows*[1] 247 com-

[1] This is similar to Twitter's notion of following.

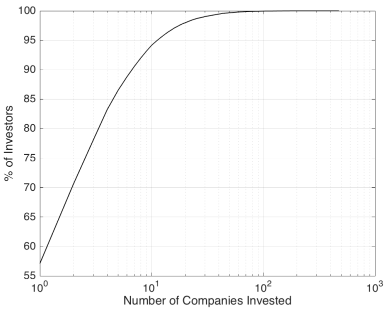


Figure 3: **CDF of number of investments made by each investor.**

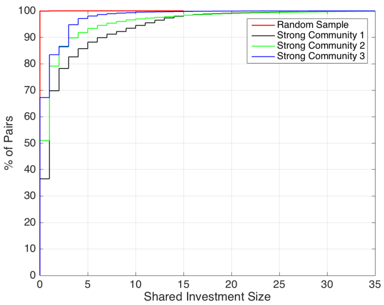


Figure 4: **Comparison of CDFs for shared investment size.**

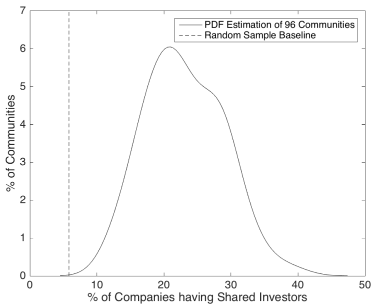


Figure 5: **PDF estimation of 96 communities.**

panies on AngelList, but makes an investment only to 3.3 companies on average, with the median being 1. The most active investor makes close to 1000 investments.

4. IMPACT OF SOCIAL ENGAGEMENTS ON FUND-RAISING

	Number of companies (%)	% Success
No social media presence	668,282 (89.81%)	0.4
Facebook only	37,762 (5.07%)	12.2
Twitter only	70,563 (9.48%)	10.2
Facebook and Twitter	32,544 (4.37%)	13.2
Presence of demo video	36,364 (4.88%)	10.4
No demo video	707,672 (95.11%)	0.9
Facebook (>652 likes)	15,510 (2.08%)	18
Twitter (>343 tweets)	32,470 (4.36%)	14.7
Twitter (>339 followers)	32,477 (4.36%)	15.2
Facebook (>652 likes) and Twitter (>339 followers)	9,944 (1.33%)	22.2
Facebook (>652 likes) and Twitter (>343 tweets)	9,685 (1.30%)	22.1

Figure 6: **Social engagement's impact on fundraising (summary table).**

Our first analysis aims to quantify the benefits of social engagements on fund-raising success. Figure 6 shows a table that summarizes our results. Based on our AngelList dataset, we categorize companies based on the presence and absence of social media websites (Facebook and Twitter) that they are using, and aim to understand the impact their engagements have on their success in fund-raising. Note that since not all companies' AngelList profile is updated, there may be companies with a social media presence but have omitted including the URL on AngelList. Hence, the presented numbers on the second column (number of companies, together with its corresponding percentage over the total number of companies) should be viewed as a lower-bound. Nonetheless, having a valid link to a social media website represents a stronger factor in fund-raising on AngelList, than having the social media presence but omitting it from AngelList. The third column shows the percentage of companies that fall in that category that has successfully raised funding (an information that can be derived from CrunchBase).

Our results show that social media engagement makes a significant difference to a startup's likelihood of success. Only 0.4% of startups without any social media presence managed to successfully raise funding. This increases significantly to 12.2% and 10.2% respectively once companies have either a Facebook or Twitter presence respectively. This represents a 30X and 26X increase of fund raising success. Interestingly, we note that having both Facebook and Twitter accounts simultaneously do not significantly increase the likelihood of funding success (13.2%), highlighting the diminishing returns of having multiple social media outlets. However, engagement on social media results in a significant boost. For example, on Facebook, companies with $>$ 652 likes (where 652 is the median number of likes across all valid Facebook accounts on AngelList), the success rate increases to 18%. Likewise, engaging users via tweets or followers on Twitter also provide a boost in funding success. Finally, AngelList allows companies to post a demo video online. We observe that companies with a demo video are at least 11.5 times more likely to succeed in fundraising.

Note that our measurement numbers are based on a snapshot of the crawled data. Hence, the observations capture correlation, not causality. For example, a company may have a more active social media presence after it has successfully raised funding (and hence have the human resources to engage in such activities). Understanding the causal relationships is an avenue for our future work (Section 7), and requires us to track companies over time.

5. INVESTOR GRAPH ANALYSIS

Our next set of analysis aims to understand the dynamics of startup investments.

5.1 Investor Graph Generation

As a starting point, from our AngelList and CrunchBase data, we extract out the list of startups invested by each investor, to identify investment patterns. The extraction is done via a parallel Spark query that merges AngelList and CrunchBase data, and then generate as output a bipartite graph connecting investors and companies they invested in.

In our crawled data set, each investor and investor has a unique AngelList identifier (ID). We extract these IDs using Spark, and then generate investment edges of the form "investor_id vs. company_id". These edges result in a bipartite graph: if each investor and company each represent a node, there will be edges from investor nodes to company nodes if there is an investment relationship between the two. Note that we omit from the investor graph generation any investors that have made no investments in the past. Then the final bipartite graph consists of 46,966 investor nodes, 59,953 company nodes, and 158,199 investment edges. On average, each company has 2.6 investors.

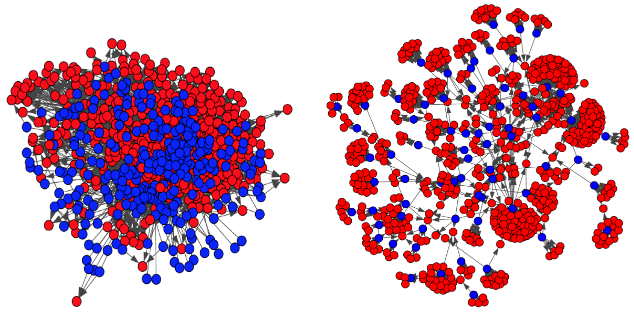
Revisiting Figure 3, we note that a small fraction of investors are actually taking a large portion of investments in the graph. This is reflected in the bipartite graph we generated. Only 30% of the investors have out-degree (number of companies invested) ≥ 3. However, these investment edges account for 75% of all the investment edges. Likewise, 22.2% of the investors have out-degree ≥ 4 but account for 68.3% of all investments. Finally, only 17.0% of the investors have out-degree ≥ 5, accounting for 62.0% of all investments.

5.2 Community Detection

Using the bipartite graph as a basis, we next aim to run community detection algorithms to cluster investors that tend to invest in similar companies together. As an initial cleaning step to make the cluster statistically meaningful, we consider only investors that have invested in at least 4 companies. We next apply the CoDA [19] community detection algorithm from the SNAP [11] library. We selected this tool since it is suited for handling bipartite graphs with directed edges.

Using the CoDA method of Stanford's SNAP library, we are able to group investors into 96 communities with an average size of 190.2. These communities reflect the tendency of investors to follow the investment patterns of others. To give a sneak preview of our results, Figure 7 shows a visualization (obtained using the Python library igraph [10]) of a strong and weak community respectively, as generated by CoDA. We will revisit metrics for actually evaluating these communities later in this section.

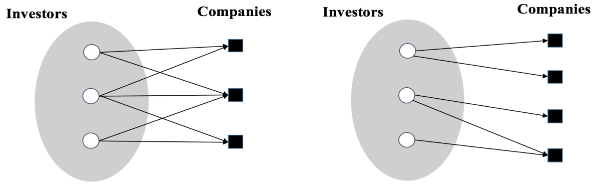
Interestingly in Figure 7a, we observe a strong community where there is significant herd mentality: many investors (blue) are co-investing in several similar companies (blue). Alternatively, Figure 7b shows a weaker community, where each investor (blue) tends to invest in its own set of companies (red) independent of other investors.



(a) **Strong community.** (b) **Weak community.**

Figure 7: **Visualization of example communities (blue: investors; red: companies).**

5.3 Community Detection Metrics



(a) **Example community 1.** (b) **Example community 2.**

Figure 8: **Toy examples for communities of investors.**

Based on the communities detected by CoDA, we next aim to quantify the strength of each community using two metrics. As our visualizations suggest, a community is considered strong if investors within the community invests in many overlapping companies. We illustrate this using a simpler example. Figure 8a shows an example bipartite for a strong community of investors, while Figure 8b shows a relatively weaker one.

To measure the strength of communities in this context, we propose two metrics. The *shared investment size* metric is defined from the investor's perspective. It counts the intersection size of two investors' investing companies sets. Suppose there are two investors: investor 1 and investor 2, and the companies sets they invest are C_1 and C_2 respectively, then their "shared investment size" is $|C_1 \cap C_2|$. For a given community, we can hence gain a measure of the strength of the community by taking the average across all shared investment sizes between all pairs of investors within the community. Using our above example to illustrate, the average shared investment size of Figure 8a and 8b are $(2+2+1)/3 = 1.67$ and $(1+0+0)/3 = 0.33$ respectively.

Our second metric is derived from the view of companies instead of investors. Within each community, we compute the *percentage* of companies with *shared investors* of at least K. For example, when $K = 2$, we identify companies that are co-invested by at least two investors from the same community, and then we compute the percentage of these companies as a percentage over all companies invested by the community. The larger the value, the most likely a company will be invested by investors from a given community. In Figure 8a, for a given $K = 2$, the percentage of companies having shared investors is $3/3 \times 100\% = 100\%$, while in Figure 8b, the percentage is only $1/4 \times 100\% = 25\%$.

To evaluate our identified communities based on these metrics, Figure 4 shows four CDFs for the first *shared investment size* metric. We select three strong communities, and compare the results against an estimated CDF across the entire bipartite graph. To estimate the CDF $F(x)$ of the uniform distribution over all the data, we pick 800,000 i.i.d. sample pairs of investors, and get the empirical CDF $F_n(x)(n = 800,000)$. By the Glivenko-Cantelli theorem [6], we can guarantee that the probability that $||F_n - F||_\infty \leq 0.0196$ is at least 99%.

We make the following observations. First, there is significant herd mentality among the top three strong communities. In the strongest community, some investors may share up to 48 co-invested companies. On average, the two strongest communities have shared investment sizes of 2.1 and 1.6, averaged across all pairs of investors in the community. This is a significantly high number, given that on average, each company has 2.6 investors, a figure that is likely tied to a cap on each company's funding goals.

On the other hand, applying the second metric from the view of companies, we compute the percentage of companies that have at least two common investors for each of the 96 communities. Figure 5 shows a PDF of the average percentages across all 96 communities. We again observe that in a number of these communities, there can be upwards of 20% of companies being co-invested by at least two investors. The average percentage across all communities is 23.1%. As a point of comparison with a randomized community of investors, we observe that the shared investment percentage is only 5.8%, which is significantly lower. This suggests again the likelihood of herd mentality among investors, though a detailed longitudinal study is required to validate this trend.

Revisiting Figure 7, the strong community has an average *shared investment size* of 2.1 and a *percentage* of companies with *shared investors* of 27.9%, while the weak community has an average *shared investment size* of 0.018 and a *percentage* of companies with *shared investors* of 12.5%.

6. RELATED WORK

Crowdfunding. Prior studies on crowdfunding have explored investor recommendations [14] based on *Kickstarter* [9], a crowdfunding site for creative projects. [18] applies machine learning and text mining techniques on news article to predict the likelihood a company can be acquired. Our work is significantly different, as we focus on the AngelList platform, which is not only more recent, but also focuses on crowdfunding for startup investments, which has very different dynamics as compared to crowdfunding for specific projects. [17] does an exploratory study to identify factors for crowdfunding success, but provides only basic analysis on macro-level statistics. Our work is significantly more comprehensive as we integrate data from a range of data sources. The focus of our work is also different from all of our prior work, given our focus on the impact of social engagements and community detection.

Community detection. Community detection also plays a significant role in our work, since we use it to study the investment decision behavior of investors. As a starting point, we use the CoDA [19] method from the SNAP [11] library. Existing community detection algorithms are predominantly focused on densely connected nodes in undirected graphs, while CoDA [19] provides us with an effective method to do community detection on directed 2-mode (bipartite) networks, which is a suitable tool for our dataset. While CoDa's algorithm appears to be effective, it does not map directly into the metrics required to quantify investment communities. As future work, we plan to come up with new community detection algorithms that are more suited to our metrics and data.

7. DISCUSSION

In this paper, we present a platform for exploratory data collection and analysis of social media websites. We focus on crowdfunding sites as our driving example. Our results in this paper are promising. We have shown that there is a strong correlation between the level of social media engagement and actual fund raising success, and in fact, fundraising success appears to be significantly higher for companies that are actively engaging users on social media websites. We further apply community detection algorithms, visualizations, and novel metrics to measure co-investment strength among investors.

Moving forward, we are extending our work along several dimensions, in the areas of social psychology and economic models for startups, and advancements in big data research to meet our meets. We briefly describe two of our ongoing work in the latter.

Causality analysis. Our one-time data collection provides a basis for correlation studies. However, this is insufficient for us to determine causality. We plan to capture a longitudinal study of social media data pertaining to startups. For example, using the AngelList example above, we can identify startups that are attempting to raise money on AngelList. We will then set up a daily data collection task that determines which startups are currently fundraising on AngelList, and using various API calls, we will gather the latest information related to their new tweets, Facebook posts, increases in likes and followers, profile updates, and press releases. As companies on AngelList start fundraising campaigns, we will determine how much money they have raised over time and the duration required by each company to reach its target funding goals. Causality analysis may be conducted to determine whether social medial engagement directly impacts fundraising success.

Community detection on dynamic graphs. By analyzing social networks over a period of time, we plan to statistically analyze correlations between the activities and a successful fundraising event. Since the group structure of the graph is not known a priori, we will use statistical methods in social network analysis to infer the community structure of the graph. We will perform community inference using stochastic block models [15], which outputs an assignment of nodes to communities based on the adjacency matrix of the graph. Since Twitter graphs are inherently directed in nature, we will investigate methods to extend stochastic block model inference procedures to directed graphs. Based on the inferred community structure, we can measure simple summary statistics of the amount of co-investing occurring in the graph.

We also plan to understand the dynamics in terms of formation or disbanding of community clusters over time. We further plan to use characteristics such as node degree, connectivity, and measures of centrality in each of the graphs in our database to predict the success or failure of a startup. Our hypothesis is that graph characteristics such as centrality will be more useful for predicting the success in the case of the Twitter graphs, since a high measure of centrality would indicate the ability of a firm to bridge investors to potential customers; whereas characteristics such as node degree would be more indicative of success in the CrunchBase and AngelList networks. We will use feature selection methods for high-dimensional regression to identify the graph statistics that are the most useful for performing prediction.

Acknowledgments. We would like to thank Sun-Young Park, Vin Marshall and Stuart Wagner for several helpful discussions on this work. This work is funded in part by NSF CAREER CNS-0845552 and CNS-1117052.

8. REFERENCES

[1] Angellist. https://angel.co/.

[2] Angellist wikipedia. https://en.wikipedia.org/wiki/AngelList.

[3] Crunchbase. https://www.crunchbase.com/.

[4] Equitynet. https://www.equitynet.com/.

[5] Fundable. https://www.fundable.com/.

[6] Glivenko-cantelli theorem wikipedia. https://en.wikipedia.org/wiki/Glivenko-Cantelli_theorem.

[7] Graph api. https://developers.facebook.com/docs/graph-api.

[8] Hdfs. https://hadoop.apache.org/.

[9] Kickstarter. https://www.kickstarter.com/.

[10] Python library igraph. http://igraph.org/python/.

[11] Snap. http://snap.stanford.edu/.

[12] Spark. http://spark.apache.org/.

[13] Tweepy. http://www.tweepy.org/.

[14] J. An, D. Quercia, and J. Crowcroft. Recommending investors for crowdfunding projects. In *WWW*, pages 261–270. ACM, 2014.

[15] D. S. Choi, P. J. Wolfe, and E. M. Airoldi. Stochastic blockmodels with a growing number of classes. *Biometrika*, page asr053, 2012.

[16] T. J. Kane. The importance of startups in job creation and job destruction. *Available at SSRN 1646934*, 2010.

[17] E. Mollick. The dynamics of crowdfunding: An exploratory study. *Journal of business venturing*, 29(1):1–16, 2014.

[18] G. Xiang, Z. Zheng, M. Wen, J. I. Hong, C. P. Rosé, and C. Liu. A supervised approach to predict company acquisition with factual and topic features using profiles and news articles on techcrunch. In *ICWSM*, 2012.

[19] J. Yang, J. McAuley, and J. Leskovec. Detecting cohesive and 2-mode communities indirected and undirected networks. In *WSDM*, pages 323–332. ACM, 2014.