Venture Capital Communities [1]

Amit Bubna
Indian School of Business
Gachibowli, Hyderabad, India 500 032

Sanjiv R. Das
Leavey School of Business
Santa Clara University, CA 95053

Nagpurnanand Prabhala
Robert H. Smith School of Business
University of Maryland, College Park, MD 20742

February 27, 2014

[1]Comments welcome. We thank Alexandre Baptista, David Feldman, Jiekun Huang, Ozgur Ince, Vladimir Ivanov, Pete Kyle, Josh Lerner, Laura Lindsey, Robert Marquez, Vojislav Maksimovic, Manju Puri, Krishna Ramaswamy, Rajdeep Singh, Richard Smith, Anjan Thakor, Susan Woodward, Bernard Yeung, and participants at the CAF, FIRS, Midwest Finance Association, World Private Equity and TAPMI conferences, and seminar participants at Blackrock, Florida, George Washington University, Georgia State, Georgia Tech, ISB, Kellogg, Maryland, NUS, the R User Group, Rutgers, UNSW, and Villanova for helpful comments. The authors may be reached at their respective email addresses: `amit_bubna@isb.edu`, `srdas@scu.edu`, and `prabhala@umd.edu`.

Abstract

Venture Capital Communities

While it is well-known that syndication is extensively used in venture capital (VC) financing, less is known about the *composition* of VC syndicates. We present new evidence on this issue. While VC firms have a large pool of syndicate partners to choose from, they tend to draw from smaller groups of partners that we call VC "communities." We implement new techniques to uncover these groups and use them to understand preferences driving syndicate partner selection. We find a complex pattern in which preferences for dissimilar partners to extend influence coexist with preferences for similarity in terms of functional style on dimensions of industry, stage, and geographic specialization. The spatial loci of community clusters suggest that syndicates compete through differentiation and specialization rather than generalized skills relevant to young firm financing. Community backed ventures are more likely to exit successfully. Our results are consistent with learning-by-doing or incomplete contracting models of VC investing in which familiarity aids learning and enhances trust and reciprocity.

JEL classification: G20, G24

Key words Venture capital, syndication, community detection, social networks, boundaries of the firm

1 Introduction

Venture capitalists raise capital from wealthy individuals and institutional investors and invest in young firms that promise high upside. According to the National Venture Capital Association, there were over 56,000 VC deals for $429 billion in the U.S. between 1995 and 2009. VC successes include many high tech firms such as Apple, Cisco, and Microsoft. See Da Rin, Hellmann, and Puri (2012) for a recent survey of VC research.

VC investing is risky. Firms financed by VCs tend to be young, and often have unproven business models. VC investing is also resource intensive, demanding considerable effort in terms of ex-ante screening and follow on support such as strategic advising and recruitment of key personnel for the portfolio firms.[1] VCs manage the risk and resource demands of investing through multiple strategies. One is the use of security design methods such as priority, staging, or contracting over control rights.[2] A key element of such strategies is syndication, or co-investing in portfolio firms together with other venture capital investors.

Syndicated deals are economically important. In the U.S., they account for 66% of VC investment proceeds and 44% of the number of rounds financed. Only 5% of VCs never syndicate and these are small, peripheral firms. While the importance of syndication is long recognized in the VC literature (e.g., Bygrave and Timmons, 1992), less is known about the preferences that drive syndicate partner selection. Do VCs pick syndicate partners at random? Or, do they draw from subsets of preferred partners? If so, what types of partners do they prefer? Are preferred partners observably similar or dissimilar, and on what dimensions are they so? We develop new evidence on these questions, and in doing so, shed light on the *composition* of VC syndicates.

We show that while VCs have a large pool of syndicate partners to choose from, they are likely to draw from smaller groups of preferred partners. Equivalently, VCs cluster into small groups of preferred-partner "communities," whose members are more likely to syndicate with each other than with others. We identify preferred-partner communities from

[1] See Gorman and Sahlman (1989), Gompers and Lerner (2001), or Hellmann and Puri (2002).

[2] See Cornelli and Yosha (2003) or Neher (1999) on security design. Kaplan and Stromberg (2003, 2004), Robinson and Stuart (2007) and Robinson and Sensoy (2011) discuss VC contracts.

observed VC syndication data using techniques recently developed in the physical sciences literature (Fortunato, 2009). We use the results to understand the structure of preferences that drive syndicate partner choices and competition between VC syndicates. Our results indicate that VC preferences are complex, with preferences for similarity on dimensions of functional style and preferences for the dissimilar on dimensions of size and influence. We find that different VC clusters represent different pools of expertise, consistent with syndicates competing through differentiation and specialization in such dimensions such as knowledge of industry and local geographic markets.

The starting point for our study is the observation that while VCs have wide choices for syndicate partners, they exhibit strong preferences for some partners over others. To illustrate this preference, Figure 1 displays the frequency distribution of the syndicate partners of J. P. Morgan between 1980 and 1999. The long and thin right tail indicates that it has many partners – over 600 in the sample period – but a thick left mass shows its preference for some partners over others. Examples of its preferred partners include Kleiner Perkins, Oak Investment Partners, and the Mayfield Fund. Figure 2 shows similar patterns for Matrix Partners, Sequoia Capital, and Kleiner Perkins.

Anecdotal evidence also suggests that VCs prefer to syndicate with a few familiar partners. Fred Wilson of Union Square Ventures, a prominent investor in major social network sites such as Tumblr, Twitter and Zynga, says "... there are probably five or ten VCs who I have worked with frequently in my career and I know very well and love to work with. It's not hard to figure out who they are..."[3] A familiar VC can be a past syndication partner, or alternatively, a VC who has dealt with a syndication partner. As Matrix Partners writes in its website, "... the best way to get in touch with our team is through an introduction from someone you know in our network."[4] If VCs prefer familiar VCs as partners, we should observe that they cluster into small groups or communities whose members prefer to syndicate with each other but not exclusively so.

Economic theory also motivates a preference for small sets of familiar partners. One mo-

[3] http://www.avc.com/a_vc/2009/03/coinvestors.html.

[4] http://matrixpartners.com/site/about_partnering-with-matrix, accessed May 3, 2011. The founder of LinkedIn, Reid Hoffman, makes a similar point in his autobiography, *The Startup of You*.

tivation comes from learning-by-doing models of VC investing (Goldfarb, Kirsch, and Miller (2007), Sorensen (2008)). VC investing is skill-intensive. While some skills are endowed, others are acquired through learning-by-doing because VC-funded firms tend to have unproven business models. Syndicating with familiar partners can aid learning through better understanding of partners' norms and processes (Gertler (1995); Porter (2000)). Incomplete contracting theories also generate a preference for familiar partners. In models such as Grossman and Hart (1986) or Hart and Moore (1990), the suspicion that partners will free ride or hold up initial investors lowers investment. These problems are alleviated when partners know each other. Familiarity can also lead to better outcomes by enhancing trust and reciprocity (Guiso, Sapienza, and Zingales (2004), Bottazzi, Da Rin and Hellmann (2011)).

Work in sociology also motivates a preference for a limited number of syndicate partners. For instance, following Granovetter (1985), agents place more weight on information flows from familiar sources. A related literature discusses the tradeoffs between familiar and unfamiliar partners. Relationships with familiar partners generate social capital (Gulati, 1995) but repeating the same set of relationships with no change also precludes access to valuable source of heterogeneous information, losing the strength of weak ties (Granovetter, 1973). Uzzi (1997) argues that something in between where agents have both strong and weak ties is likely optimal. This description is exactly the intuition for the preferred-partner communities we detect and study. VCs belonging to a community predominantly partner with their community members. However, nothing precludes them from less frequently partnering outside, as we find in the data. Section 2 formalizes this intuition, develops the mathematical representation of groups with simultaneous strong and weak ties, and discusses the optimization problem we solve to detect these groups.

Our first results identify preferred-partner communities. We comment on the technique as it is new to the finance literature. Our methods take as raw data the history of partnerships between VCs in past syndications. From this data, we identify communities, or small clusters of VCs with a high propensity to do business with each other. Community detection is thus a clustering technique. It is, however, different from clustering methods used in finance, e.g., Brown and Goetzmann (1997). We briefly consider the differences.

Standard clustering problems optimize variation within relative to variation across clusters. Here, we optimize *syndication probability* within relative to outside. A second difference is the considerable flexibility we allow in cluster formation. We do not pre-specify the number of clusters or their size. We allow for multiple clusters of different sizes. We do not require that all VCs should belong to clusters. We leave cluster boundaries open to let members syndicate within and outside their preferred partner groups. The flexibility comes at a price of considerably greater computational complexity. Exact solutions to the problem of detecting such flexible clusters are not known but algorithms to solve the computational problem have developed over the last decade (Fortunato, 2009).

Our data comprise U.S. VC syndications between 1980 and 1999. Empirically, we detect communities in every 5-year period of our sample. About 20% of VC firms in each period belong to communities and the median community size is 13. We quantify the strength of ties within to ties outside. On average, a community VC is 16 times as likely to syndicate within the community than outside. We test whether communities are stable by computing the Jaccard index, or the fraction of common members shared by all overlapping communities in adjacent periods. As the statistical distribution of the index is not known, we follow the event study literature (Brown and Warner, 1985; Barber and Lyon, 1997) and establish significance through simulations. Our communities are far more stable than under the null. Thus, we find a large number of preferred-partner groups among VCs. These groups are both tight-knit and stable.

We next consider the characteristics of a VC's preferred partners. One view is that a VC's preferred partners should be similar. The behavior could arise because of an underlying behavioral trait, the "birds of a feather flock together" viewpoint of McPherson, Smith-Lovin and Cook (2001). Theories of contracting with private and manipulable signals (Cestone, Lerner and White (2006)) also predict that preferred partners should be similar. In these models, one syndicate member's optimal quality is increasing in the other's quality. Finally, VCs may place more faith in partner judgments if the partners are functionally similar. Each of these forces predicts that preferred-partner communities contains members who are relatively similar to each other.

A contrasting view is that VCs prefer dissimilar syndicate partners. As Hochberg, Lindsey, and Westerfield (2012) discuss, preference for heterogeneity could arise because heterogeneous partners give VCs access to broader skill sets, resources to help portfolio firms, or greater reach in new domains. Preferences for different partners could also arise due to complementarity-seeking behavior. Here, partners strong on one dimension but weak on another can seek partners with strength in the dimensions they are weak on, resulting in preferred partner groups with higher variation along both dimensions. For instance, in the biotechnology industry, firms strong in drug development often partner with larger firms with skills in post-development marketing and manufacturing (Robinson and Stuart, 2007).

Tests of whether VCs prefer similar or dissimilar partners require specification of a set of observable attributes along which there is similarity or dissimilarity seeking behavior. While our choices largely reflect prior VC literature, our x-variables fall into two broad categories. One reflects a VC's functional style and the second set measures a VC's influence or reach. We also need to assess the significance of the similarity or dissimilarity between members in clusters of varying size and number in each period. As the analytic distributions of the test statistics are unknown, we generate null distributions through simulations.

The test results suggest that partner preferences are subtle, encompassing both preferences for similarity and preferences for dissimilarity along different dimensions. The dissimilarity between preferred partners is primarily along dimensions of influence, consistent with partner-seeking behavior to extend a VC's reach. Similarity-seeking behavior is along dimensions of functional style. Preferred partners are similar in terms of how they spread capital within stage, industry and portfolio geographic location, and also in terms of the specific stage, industry, or geographic location based expertise. The result is a mix of VC disassortativity on dimensions of influence coupled with assortativity on dimensions of style.

We next consider the spatial loci of different preferred partner communities. One view of preferred partner groups is that they are soft conglomerates that house a broad vector of skills to cater to a broad range of businesses. Alternatively, different community clusters could locate in separate portions of the style space so that each offers a specialized set of skills. We test these two views by examining the distances between communities along different

attribute dimensions. Our evidence is more consistent with the differentiation view in which different clusters represent different specializations on dimensions such as the knowledge of local geographic markets or specific sectors.

Finally, we investigate the performance of firms funded by community VCs. A key issue in all large-scale VC studies is that exits via M&As are noisy, reflecting a mix between failures and success stories such as Skype's acquisition by Microsoft. We follow the best practices in the VC literature by considering multiple definitions of exit, including the most stringent one, an IPO, and a competing risks specification in which M&As and IPOs are alternative forms of exit in a competing hazards duration model. There is a positive relation between successful exit and getting funding from a community VC rather than non-community VC, a result robust to falsification placebo tests.

The rest of the paper is organized as follows. Section 2 formally defines communities with strong internal and weak external ties, develops the related optimization problem and discusses solution methods. Section 3 discusses data. Sections 4 and 5 present the main results and robustness tests. Section 6 concludes and suggests directions for future research.

2 Preferred-Partner Communities

2.1 Definition

Given a set of VCs, it is straightforward to define sets of preferred partner communities. These are clusters of VCs with the property that the members of each cluster have strong syndication ties with each other than with non-cluster members. We do not require that all VCs exhibit such behavior so not all VCs necessarily belong to preferred partner groups.

2.2 Mathematical Representation

We partition the space of all VCs into K clusters. We suppress time period subscripts t for compactness. There are K preferred-partner clusters with $n_k \geq 2$ members per cluster, $k = 1, 2, 3, ...K$. We have $N_k = \sum_{k=1}^{K} n_k \leq N$ where N is the number of VCs so there are

$N - N_k$ singleton VCs who are not cluster members. We derive K and n_k endogenously as an output of the clustering process instead of imposing arbitrary constraints.

With this background, we specify the community detection problem. Formally, let c_k denote community k and $\delta_{ij}(c_k)$ be an indicator variable equal to 1 when both VC i and VC j belong to community k. Define the *propensity to syndicate* within a given cluster as the actual number of syndications within a cluster minus what is expected by chance. Modularity Q is defined as the sum of in-cluster syndication propensities across all clusters. The mathematical problem of community detection is to choose the optimal number of clusters K, the size of each cluster n_k, and the cluster membership, i.e., the set of indicator variables $\delta_{ij}(c_k)$, to maximize modularity Q, where

$$Q = \frac{1}{2m} \sum_{k} \sum_{i,j} \left[a_{ij} - \frac{d_i \times d_j}{2m} \right] \cdot \delta_{ij}(c_k) \tag{1}$$

where $d_i = \sum_j a_{ij}$ is the number of syndications done by VC firm i (j) and $m = \frac{1}{2} \sum_{ij} a_{ij}$, with the factor of 2 to reflect the equality of ties between i and j and ties between j and i.

The first term in the square bracket in Eq. (1) represents the actual number of deals co-syndicated by VCs i and j. The second term in [...] represents deals expected to be co-syndicated between i and j purely by chance. Intuitively, VC i with many connections will have greater odds of syndication with *any* VC j. Thus, the numerator in the second term $d_i \times d_j$ is increasing in the number of connections of VC i. The difference between the two terms represents cluster k's propensity to in-syndicate.

Modularity Q sums the in-cluster syndication propensities across all clusters c_k. Q lies in [-1, +1]. $Q > 0$ means that intra-community ties exceed ties predicted by chance. There are no known exact solutions for Q-optimization beyond tiny systems. The problem is computationally intensive given the large number of feasible partitions given the flexibility in number of clusters, cluster sizes, and because we permit open cluster boundaries.[5]

Fast solution methods include agglomerative techniques that start by assuming all nodes are separate communities and build up clusters iteratively. For instance, the particular

[5]See Fortunato (2009) for a discussion of these issues and solution techniques.

technique we use, the walk trap method, initiates simultaneous random walks at several nodes, each taking a step in a random direction. Communities are sets of VCs from which the random walks fail to exit within a fixed number of steps. Intuitively, if a set of VCs are tight-knit with a high propensity to syndicate with each other, random walks initiated with any one VC are likely to spend longer periods of time in the clusters. Appendix A gives R code for implementing the algorithm based on Pons and Latapy (2005).

2.3 Relation to Social Networks Literature

Our study is related to work on social networks in finance. One research stream examines the pairwise connections between agents, such as ties between CEOs and directors, or directors and analysts. Another stream examines aggregates derived from the pairwise ties. We discuss this work and position our study relative to both strands of the literature.

The starting point in the networks literature in finance is the pairwise connections between agents. For instance, the pioneering studies of Cohen, Frazzini and Malloy (2010, 2012) examine ties between analysts and boards of directors of firms they cover. Hwang and Kim (2009) and Chidambaran, Kedia, and Prabhala (2010) analyze links between CEOs and directors. In the VC context, Bhagwat (2011) and Gompers, Mukharlyamov, and Xuan (2012) examine connections between executives employed at VC firms based on VC executive biographies. Connections between VC firms, founders, and top executives are studied by Bengtsson and Hsu (2010) and Hegde and Tumlinson (2011). Taking these pairwise ties as x variables, the studies explain dependent variables such as information flows, fraud or exit.

A second stream of research does not stop at pairwise ties but aggregates the ties of an individual agent to compute aggregate connectedness metrics.[6] For example, the influence of an individual VC, i.e., centrality, is determined by the number of her syndicate partners and in turn their connections. Hochberg, Ljungqvist and Lu (2007) introduce centrality to the finance literature. They find that central VCs are more successful ex-post. Engelberg, Gao and Parsons (2010) find that central CEOs are paid more. Stanfield (2013) studies the

[6]For a textbook treatment, see http://faculty.ucr.edu/ hanneman/nettext/

role of LBO sponsor centrality in predicting LBO exits.

The key metric in our study is community membership. Like centrality, it is an aggregate derived from pairwise ties but the two measures have little else in common either operationally or economically. Operationally, centrality is a raw or weighted sum of an agent's connections while community membership is a solution to the problem of optimizing Q in Eq. (1). Economically, the two measures capture very different economic intuitions. Centrality counts the number of ties of an agent. Community membership identifies entire *groups of agents* who tend to do business together. As an illustration with a real-life example in finance, consider Figure 3, reproduced from Burdick et al. (2011). The three banks with high centrality are Citigroup, J. P. Morgan, and Bank of America. None belongs to communities, which are in the left and right nodes of Figure 3.

To further illustrate the differences between community and centrality, consider the VC context. Following Hochberg, Ljungqvist and Lu (2007), a high centrality VC has worked with many partners. However, centrality says very little about the identity of a VC's preferred partners. Centrality sheds no light on whether a VC has a diffuse set of partners or prefers some partners over others. It does not say which partners are preferred or what attributes the preferred partners have. However, these types of questions are at the heart of our study of communities. Neither community nor centrality implies the other nor is one a proper subset of the other. One is about clusters, or groups who work together in syndicate deals, while the other is about a single VC's reach.[7]

2.4 Relation to Pairwise Ties Literature

Our study extracts entire groups of VCs who show a propensity to syndicate with each other. Papers such as Du (2011) and Hochberg, Lindsey, and Westerfield (2012) study pairwise models of VC tie formation. An interesting and open question is how we can reconcile the two approaches. In our view, the two techniques are complementary in a sense that we

[7]In fact, the physical sciences literature on communities barely refers to centrality. For further discussion of these distinctions, see Sections 7.1, 7.2, and 11.6 of Newman (2010). Other applications of community detection include identifying politicians who vote together (Porter et al., 2007). product word groups (Hoberg and Phillips, 2010), collaboration networks (Newman, 2001). Others are discussed in Fortunato (2009).

elaborate upon next.

Underlying the process of picking VC syndicate partners is a structural model that describes what partnerships form for a given financing opportunity. This is likely a complex dynamic model in which VCs choose partners based on their own prior interactions with the same partner or with other VC partners and the characteristics of the firm being financed. The error terms in this choice model have complex unknown time series and spatial correlation structures.

One approach towards understanding the true model is to actually specify it and estimate its true parameters such that it reproduces observed data, or the set of realized syndications over a period of time such as 5 years. This is a complex task as the analytics for even simpler network formation models are difficult (e.g, Currarini, Jackson, and Pin, 2012). With this background, the reduced form i.i.d logit models of pairwise tie formation such as those of Du (2011) and Hochberg et al. (2012) can be viewed as a necessary and useful first step towards developing a full structural model.

Ours is the converse top-down approach. Instead of starting with the bottom-up approach of modeling pairwise ties with all attendant complexities of the dynamics and spatial and temporal error correlations, we start with the *end product* of the tie formation process. This is the actual set of observed syndicate partnerships established by VCs over a period of time, which represents the revealed preferences of VCs in their syndicate partner choices. Our approach is to invert these observed choices to infer the drivers of VC behavioral preferences. For instance, we can test whether partner preferences are diffuse or concentrated and the attributes that drive these preferences.

3 Data

We analyze VC investments made from 1980 to 1999 in Thomson's Venture Economics database. The sample period starts in 1980, roughly corresponding to the institutionalization and growth in the VC industry (Gompers and Lerner, 2001). While the data go to 2010, our sample ends in 1999 to allow sufficient time to observe successful ex-post outcomes of

investment. The unit of observation is a round of financing by U.S.-based VC funds.

We refine the initial sample by dropping buyouts, deals in which we cannot identify investors or there are only individual investors such as angels or management. We do *not* exclude deals that involve institutions such as subsidiaries of financial institutions and technology transfer offices of universities. While these investors have different incentive schemes, there are two reasons to keep them in sample. First, the *deals* they finance involve traditional institutional VCs. Second, each syndicated deal, whether between institutional VCs or by institutional VCs with others, offers VCs an opportunity to interact and for partners to learn about each other. To the extent the conventional institutional VC firms we are interested in learn from such deals and different partner types, we include the latter in our final sample.

A related question is whether we should include only first round financings or first and subsequent round financings. Here, there are two very different conceptual questions. One is a mechanical question of miscoded data. These are easily ruled out by discarding successive rounds of financing within (say) 60 days. The economic question is whether there is any active learning in second and higher rounds of financings. If later rounds are passive and involve more routine interactions, familiarity with partners counts for less. If second and later rounds involve substantial interactions and learning, they matter as much as first rounds. With no clear guidance about the right approach, we report both sets of results. We take as our primary working assumption that all rounds involve some learning. However, the specialness of first round financings motivates us to consider round 1 and later rounds separately. We report some interesting differences.

The last part of our paper examines investing outcomes through portfolio firm exits. VC firms can exit through mergers and acquisitions (M&A's) or through IPOs. We obtain data on IPO firms from Thomson Financial's SDC Platinum. We match companies by their CUSIP identifiers, cross-check the matches against actual names, and further hand-match the names with those in the Venture Economics database. 1,470 ventures in our sample exit via IPOs. We obtain M&A data from Thomson Financial's SDC M&A database. There are 3,545 exits via mergers in our sample.

An important issue in the VC literature is whether exits via M&A are indicators of

success. Two studies delve into this question in detail using small subsets of venture backed financings. Kaplan, Sensoy, and Stromberg (2002) track 143 VC investments. For a subset of 500 companies, Maats et al. (2008) compare Venture Source and Venture Economics (commercially used data in virtually all VC studies) and cross-check with data on two funds that they privately obtain. The consensus from both studies is that IPO exits are accurate but M&A exits have greater noise, divided between failed financings and success stories.

There is no consensus on how best to deal with the noise in M&A exits in large sample VC studies such as ours because the samples with precisely measured exit are small subsamples of the whole VC financing datasets. The VC literature's response is to use (a) conservative; and (b) multiple measures of exit. The conservative approach is used recently in Gompers, Mukharlyamov, and Xuan (2012), who ignore M&A exits and define success as an IPO exit. We adopt this approach and modify it by using a competing hazards model in which an M&A is a competing hazard for an IPO. This specification refines the IPO-only definition of successful exit by recognizing that IPOs are observed only if an M&A does not precede the IPO. For robustness, we also follow Hochberg, Ljungqvist, and Lu (2007) and use the event of an M&A or an IPO and the progression to a next-round financing as indicators of successful exit.[8]

Table 1 gives descriptive statistics for our sample at the level of the VC firm. Our sample includes 1,962 unique VC firms. On average, a VC firm invests in 22 portfolio firms and 48 rounds. Each round involves investment of $1.95 million. Close to three-quarters of the deals made by a VC are syndicated and about one-third of the rounds are classified as early stage investments. The total funds raised by a VC amount to about $128 million (median = $17.5 million). The average age of each VC at the time of its last investment in our sample is a little less than 10 years. The VC headquarters are located in 127 Metropolitan Statistical Areas (MSAs) in our sample, with an average of about 14 VCs per MSA (median of 3 VCs). The two big VC clusters in California (CA) and Massachusetts (MA) account for about 35%

[8]The alternative approach of precisely coding every M&A exit is burdensome as it requires manual intervention and is likely worth a study in its own right. Exit forms only one part of our study and not its main focus, which is on partner preferences. We are in the process of parsing M&A exits and plan to incorporate the results in future draft.

of the VC firms' headquarters.[9]

4 Number of Communities and Community Stability

4.1 Number of Communities

Following Hochberg, Ljungqvist, and Lu (2007), we use overlapping windows of 5-year length to detect syndication patterns. Thus, the first window uses VC investments in all financing rounds from 1980 to 1984, the second one employs 1981-1985 investments, and so on. The windows allow sufficient time to identify preferences for syndicate partners but avoid excessively long periods that may contain stale information. We require a minimum community size of five members and require that the end-to-end diameter[10] not exceed one-fourth that of the entire network. This constraint is not binding in our dataset.

We find several communities in the time periods in our dataset. VCs cluster into between 12 and 35 communities (based on assuming that $\eta_k \geq 5$) that prefer to syndicate with each other. There is considerable variation in community membership status. About 20% of the active VCs belong to communities. The median community has 13 members. Figures 4–7 depict communities for four non-overlapping 5-year windows, viz., 1980-1984, 1985-1989, 1990-1994, and 1995-1999. We show the members of the largest three communities in different colors. The upper plots in each figure show the entire VC network. To present a less cluttered view of the network, the lower figure plots the largest community within all communities of at least 5 members.

We see greater density of connections within the largest community than its connections across communities. We also see a distinction between centrality and community membership in our sample. In Figure 4 all large communities are connected to one another, but in Figures 5 and 6 there are satellite communities that are large but disconnected from all other communities. Figure 7 shows satellite communities in the upper plot but the largest

[9]Some VCs may have satellite offices that we do not include in the current analysis. To the extent larger VC firms have such offices, the effect is picked up in proxies for VC size.

[10]Diameter is the longest of the shortest paths from one node to another within a community.

communities are well connected to the remaining communities. In the lower plot, all large communities are connected. Peripheral ones at the edge of the network are relatively isolated.

4.2 Community Stability

We examine community stability in two ways. One test examines community *status*. We ask whether a VC who belongs to a community in one period belongs to *some* community in another period. Table 2 reports the results. On average, 90% of community VCs continue to belong to a community in the next period and 75% of community VCs remain community VCs five years later. The community status of a VC is stable but not invariant.

The more difficult question is whether VCs tend to belong to the *same community* in successive periods. A single community could stay unchanged across two periods or break up into two communities whose union contains the original cluster. It could also break into multiple communities. When there are multiple communities of different sizes in the first period, each of which can break in arbitrary ways and intersect members of other communities from earlier periods, it is no longer possible to establish lineage. We cannot tell which community in period $t+1$ was "formed" from which community in period t.

We quantify community stability using the Jaccard index, which is an index of similarity between two sets. If A and B are two sets, the Jaccard index is $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$, i.e., the ratio of the number of common elements in sets A and B divided by the number of unique elements across both sets. Thus, the index measures the extent to which A and B overlap and is easily adapted to measure similarity between two communities in a period. This still leaves open the question of aggregation across all communities. For this purpose, let $A^t = A_1, A_2, ..., A_m$ be the set of m communities at time t, and $B^{t+1} = B_1, B_2, ..., B_n$ be the set of n communities at time $t+1$. For each community A_i, we determine a composite Jaccard measure, $JC(A_i, B) = \overline{J(A_i, B_j)}_j \mid J(A_i, B_j) > 0$ for all $j = 1, 2, ..., n$. That is, we consider the Jaccard index for all subsets of communities A and B in successive periods where the intersection of the two is non-null. The average measure across all initial communities is a composite Jaccard index.

Table 3 presents the composite Jaccard measure for adjacent periods of time averaged across all 5-year periods. We use simulated communities to establish significance levels. Specifically, we simulate communities with number and size distribution equal to what we find in every 5-year period in the data. We determine the average Jaccard index between communities in adjacent time periods as above. The significance tests are based on the empirical p-values of the distribution of the Jaccard index for 100 simulations.

For all periods, we find that the Jaccard measure of our community is greater than that of bootstrapped communities at the 1% level of significance. Thus, the composition of communities in our sample is far more stable than would occur by chance. The stability makes us more comfortable with the economic interpretation of communities as reflecting VC partner preferences. If partner preferences are stable, communities should exhibit some degree of stability across time periods, which is what we find.

5 VC Community Composition and Competition

In this section, we examine the composition of VC communities, to assess what types of syndicate partners VCs prefer. We also examine the spatial locations of different communities to understand the nature of competition in the VC market.

5.1 Which VCs Belong to Communities?

Table 4 reports descriptive statistics for financing rounds in our sample classified by whether rounds are financed by a community VC or a VC not in a community. In our sample, 15,220 out of 33,924 rounds (about 44%) are community rounds and these account for 66% of proceeds. 10,056 out of 14,897 syndicated rounds, or 67%, are community rounds. Early stage rounds account for about a third of the sample and 45% of these are community rounds. 16,270 deals or close to one-half of the investment rounds are in the CA/MA geographical clusters. This reflects a concentration of VC investments in these states as well as their representation in VC databases (Kaplan, Sensoy and Stromberg (2002)).

Venture Economics classifies VC firms into 10 industry categories. Panel B of Table

4 shows that the software industry accounts for the largest share of financing rounds in our sample, followed by medical or health firms, communications and media, and internet firms. Interestingly, community VC is more likely for the riskier and complex business models characteristic of software businesses and less likely for consumer product or industrial businesses. The finding indicates that VC firms draw on preferred partners more when facing greater uncertainty, consistent with Cestone, Lerner, and White (2006).

Panel C in Table 4 describes key characteristics across rounds. There is greater investment in rounds with a community VC ($5 million) than in rounds with no community VCs ($3 million). Besides higher investment per round, community rounds tend to have more VC firms than rounds with no community VC, both in syndicated and non-syndicated rounds. For instance, community VC syndicated rounds have 3.85 VCs on average compared to 2.64 VCs in rounds with no community VC. This pattern holds for early stage rounds and initial financing rounds. Panel D gives exit information, which we analyze and discuss later.

5.2 Partner Preferences: Hypotheses and Attributes

In principle, VCs could seek partners who are similar or dissimilar to themselves. Similarity-seeking behavior among VCs could reflect a behavioral preference for interacting with people of similar backgrounds, as discussed in McPherson, Smith-Lovin, and Cook (2001). Alternatively, Chung, Singh, and Lee (2000) argue that status-based homophily can prevail because it has signaling value to outsiders. Cestone, Lerner, and White (2006) develop theoretical models in which partner similarity acts in complementary ways to generate the best financing decisions. These arguments suggest assortative matching in which VCs are more likely to prefer similar VCs as syndicate partners.

The case for disassortative matching is based on the benefits of diversity. For instance, funds skilled in raising capital may partner with niche focused funds with skills in specific sectors. As Hochberg et al. (2012) point out, complementarity seeking behavior could also result in preferences for the dissimilar. For instance, if there are two attributes X and Y characterizing VCs, complementarity-seeking behavior suggests that high X, low Y VCs

should prefer as a partner low X high Y VCs. The net effect is that preferred partners have high variance in both X and Y. As a concrete example, the large VC firm Kleiner, Perkins, Caufield, and Byers largely prefers to invest in the clean tech area with the smaller VC firm Foundation Capital, which has greater domain expertise in the area. Foundation Capital benefits from the fund-raising capability and the reputation of the larger Kleiner, while Kleiner accesses niche expertise that a small firm such as Foundation Capital brings.

It is important to note that matching need not be assortative or disassortative uniformly across *all* attributes. The received VC literature makes specific predictions about dimensions of similarity. In the learning and vetting hypothesis for syndicate formation (Sorensen, 2008 or Cestone, Lerner, and White, 2006), the evaluation and screening abilities of partners drive syndicate partner preferences. If knowledge is style-specific, within-community similarity should primarily be along the dimension of functional expertise of VCs. Under this view, heterogeneity preference is limited to dimensions that proxy for a VC's radius of influence. Our specification of VC attributes follows along these lines.

Following the above discussion, we specify two categories of attributes. One category of attributes proxy for a VC's reach and influence. Following Hsu (2004), one is VC age, which is the difference between the VC's last investment in year t and the VC's founding year. The second proxy is the VC's assets under management, which is a direct measure of its dollar resources. Finally, following Hochberg, Ljungqvist, and Lu (2007), we consider eigenvector centrality, which measures influence based on a weighted sum of a VC's total number of connections. Greater centrality implies more influential VCs.

A second set of attributes reflects a VC's investing *style*, i.e., the set of specific asset classes or investment types that the VC focuses attention on. Placement memorandums used for fund raising articulate investing styles. While these descriptions are not legally binding, they have bite as they form the basis on which limited partners allocate capital. In fact, as Coller Capital's 2008 Global Capital Barometer report finds, 84% of limited partners do not look upon changes in stated styles (or style drift) favorably.

Existing literature suggests three major dimensions of VC style: industry, stage and geography. (Sorenson and Stuart, 2001; Chen, Gompers, Kovner, Lerner, 2010; Tian, 2011).

Anecdotal evidence is consistent with these dimensions of functional styles. It is common for a VC fund to identify an industry focus in the formal agreement with limited partners (Lerner, Hardymon, and Leamon, 2007). An illustration of the three dimensions of focus is the case of SV Angel, a "seed-stage" fund with "... tentacles into New York media and the advertising world."[11] We consider style based on a VC's past investments in different sectors and portfolio firm stages reported by Venture Economics. With regard to geography, we consider both the location of the portfolio companies in which the VCs invest and the location of the VC's headquarters. Investing experience in a region results in information flows about the resources in the region while the location of senior management of a VC can also identify hard and soft information about investments in a geographic area.

5.3 Test Statistics

To examine the nature of VC partner preferences across attributes, we compute the within-community variation of measurable VC attributes. We benchmark these results by comparing the variation to similar variation for simulated communities. The simulated communities are equal in number and have the same number of members as what we actually observe in the true preferred-partner communities detected in the data. Lower within-community variation relative to the null indicates a preference for similar VCs as syndicate partners.

For continuous characteristics (such as age, VC's assets under management (AUM) and centrality), the within-variation is the standard deviation of the attribute for VCs within the community, averaged across all communities. For discrete variables, we consider a style similarity measure based on the Herfindahl-Hirschman Index (HHI) by category for each VC. If VCs are functionally similar, they should have similar HHIs. The dispersion of HHIs measures the functional similarity of VCs. HHIs can be computed based on the number of transactions in each style bucket or the proceeds. We report the former but obtain similar results for the latter.[12] While HHIs vary from VC to VC, we also consider the geographic

[11] http://techcrunch.com/2011/05/24/sv-angel-partners-with-lerer-ventures-to-cross-syndicate-valleynyc-deals

[12] Both measures are reasonable. The key resource in a VC firm is partner time, which first order scales by the number of investments. On the other hand, the capital at risk is proportional to the proceeds invested

headquarter concentration of all VCs for a community. We compute the HHI based on the proportion of a community's VCs in each geographic location.

Besides HHI, we consider a second measure that incorporates both the extent of specialization of a VC *and* the specific sectors the VC specializes in. Consider, for instance, one VC with 100% focus in software and another with 100% focus in biotech. Both will have sector HHI's equal to 1.0. The community containing both VCs will (correctly) show no dispersion in HHIs as the two VCs are similarly concentrated in allocating their capital proceeds. However, the two VCs differ in the specific sectors they focus on, which is not picked up by the HHI variation. To capture the differences in the sectors receiving capital allocation, we compute standard deviation of the fraction of deals in each bucket across all VCs in a community. We average the standard deviation across all buckets within a community and then across all communities. Formally, let the fraction of assets flowing into bucket j for VC i in community k be f_{ijk}. We compute the standard deviation $\sigma_{jk} = \sum_{i=1}^{n} \frac{(f_{ijk} - \overline{f_{jk}})^2}{n-1}$. The average standard deviation across all k communities indexes similarity in functional focus.

Operationally, we use five stage variables in Venture Economics (early stage, expansion, later stage, other, and startup/seed), the 10-industry classification used in Venture Economics, and experiment with a variety of geographical clusters. We use the very granular MSA to state-based locations to the 14 region classifications (e.g., Northern California, Southern California, New England). We obtain similar results under all approaches.

5.4 Partner Preferences: Results and Discussion

Table 5 reports the average characteristics of VC community members in our sample. Community VCs are older, larger VCs with greater centrality and are more focused on industry, stage, and geography compared to simulated communities. The results add to the descriptive information in Table 4 and supplement it with p-values based on simulations.

Table 6 tests hypotheses about within-community similarity and dissimilarity. We report the within-community variation measures for various attributes for both observed commu-

in a firm.

nities and simulated communities, and the p-value based on simulated communities. Panel A reports the results for attributes that proxy for VC influence and reach. While preferred partner groups appear to be more homogeneous in terms of age, they appear to be more diverse or have greater variation, in terms of two measures of influence, assets under management (AUM) and centrality. On both dimensions, there is greater variation within observed communities than in simulated communities. The differences are economically significant and statistically significant at the 1% level.

Panel B in Table 6 focuses on attributes relating to functional style. We find clear evidence of homogeneity along several dimensions of style. The first three rows of the Panel show that communities tend to have lower variation in HHIs than simulated communities. Thus, generalist VCs prefer as partners other generalist VCs, while concentrated VCs prefer to syndicate with other concentrated VCs.

The rows below the HHI statistics in Panel B report variation within communities by percentages invested in each industry, each stage, or each geographical area of portfolio companies, respectively. Here, the bar for similarity is higher. A low variation requires not only a pairing of generalists (specialists) with other generalists (specialists) but also matching on the specific areas of functional expertise. For instance, low variation in industry implies that focused VCs prefer as partners other focused VCs *and* that both have relatively similar distributions of deals across specific industry sectors.

Panels C and D focus on the location and ownership of VC firms. Based on the proportion of a community's VCs in different geographic locations, we calculate the location HHI, and present the average HHI across all communities. We again find evidence of homogeneity wherein communities draw VCs from similar geographies. Communities have greater geographic HHI than simulated communities. Similarly, we find that VCs in any community have similar ownership form.

The results in Panels B-D provide strong evidence of style homogeneity in partner preferences. We find that preferred partners are similar in stage and industry preferences as well as ownership and geographic preferences at several levels of granularity. While our preferred geographic division is the 14-region classification given by Venture Economics which reflects

relatively homogeneous clusters of operation of VCs, we also report results for the more granular MSA classifications and the less granular level of the state. In all cases the results are similar: VCs prefer as partners similar functional style VCs. These results provide strong support for the view that syndicate partners perform a vetting function. The prediction of these models is that functionally similar VCs should prefer to partner with each other in syndicates, as pointed out by Cestone, Lerner, and White (2006). We find exactly this pattern in the data.

One concern about our analysis based on the share of deals in each bucket may be about the large number of empty or sparsely populated style cells. While this concern is largely mitigated by our practice of benchmarking relative to simulated communities, we also consider an alternative approach that focuses on well populated cells. In each time period of 5 years, we consider within-community variation in the fraction of deals in the top 2 stages, top 3 geographic areas of portfolio companies, and the top 4 industries identified in each sub period. Table 7 reports similar results for variations within each bucket separately, for industry, stage and geographic region. VCs within communities exhibit lower variation in each of the separate buckets of industry, stage and geographic region, providing evidence of similarity among community VCs in terms of functional expertise as well.[13]

In sum, we find that VCs are not averse to reaching out to dissimilar VCs such as ones who are differently networked from themselves, provided there is similarity in functional style. Thus, there are elements of both assortative and disassortative matching in syndicate partner preferences. The results make a broader point about "diversity" as an empirical construct. Our results support those in Hochberg et al. (2012) and we join them in emphasizing that there is no particular reason why teams should be uniformly similar or dissimilar along all dimensions. Understanding the dimensions on which there is similarity and those in which there is dissimilarity is perhaps more informative and useful than attempting to force fit all attributes to generate a single diversity index (Harrison and Klein, 2007).[14]

[13]The top industries, stages and geography of interest change over time. For instance, consumer products is in the top-4 in the early 1980s, but Internet industry replaces it after the 1990s. We also consider cosine similarity measures as in Hoberg and Phillips (2010) rather than our variation measures and obtain similar results.

[14]Having said this, we also examine cosine similarity of VCs to judge the aggregate similarity between

5.5 Competition Between VC Syndicates

In this section, we examine the nature of competition between syndicates in the venture financing market. Communities represent sets of VC firms who tend to syndicate with each other. While the previous analysis focused on differences *within* communities, this section tests the differences *between* communities to assess competition between VC syndicates.

One view of competition between syndicates is that syndicates differentiate by offering specialized skills. For instance, specialist knowledge may be required to finance clean energy due to technical expertise or knowledge of sector-specific regulations. Special expertise may be needed for assessing novel therapeutic protocols for cancer treatments. The opinions and judgments of multiple VCs specializing in the same sector become useful and VCs should seek preferred partners in the same skill area. This model of differentiation suggests that communities of preferred partners choose different spatial locations. The differences *between* communities are complementary to similarity *within* communities.

The alternative viewpoint is that communities are similar to each other, effectively acting as "soft" conglomerates that pool a broad vector of skill sets to portfolio firms. "Generalist" communities could arise if there is generalized management skill, as in Lucas (1978) or Maksimovic and Phillips (2002), that is important to all forms of VC investing. For instance, early stage firms may have good ideas within specific functional domains but may lack the organizational, management, or financial expertise to scale the ideas and translate them into successful businesses. If this type of skill is important and scalable across a broad range of firms seeking venture financing, we should observe communities choosing similar locations in the VC attribute space. Functional style similarity *within* communities a similar distribution of capital allocation by VCs within a community. The across differences capture whether communities in aggregate separate themselves through differentiation or not.

Our empirical strategy is to identify the centroid of communities along each style dimension. We test whether the style distance between community centroids is greater than what we observe for simulated communities. Greater distance implies specialization and

VCs across all attributes within a community and find a preference for the similar.

differentiation while spatial proximity along a style dimension implies that it is not a basis for differentiation. Table 8 reports the results. We find evidence of specialization along all three style dimensions of stage, industry, and location. There is less support for the view that communities are conglomerates that pool skills and compete with each other . Rather, different communities appear to be source of different sets of specialized VC syndicates.

5.6 Performance

Our last tests examine whether sourcing financing from a community VC is associated with quicker exit. Given the noise in properly classifying M&A exits as successes or failures, as discussed in Section 3, we consider multiple measures of successful exit.

We precede our discussion of the results and specifications with two comments. First, we do not estimate causal effects (Roberts and Whited, 2011).[15] Moreover, the theoretical arguments for communities suggest that they act through better ex-ante selection rather than ex-post casual effects, so the effects of our x variable do not necessarily rely on causality. Thus, we do not focus on disentangling causal effects. We focus instead on establishing robustness of the performance results through placebo falsification tests.

Panel D of Table 4 presents univariate performance results. We find that 12,622 (or 37%) of financing rounds exit through IPOs or M&As. IPOs account for 11%, or a third of these. In community rounds, 14% exit through IPOs and 29% exit through mergers compared to 9% and 24% for non-community VC rounds, respectively. We find similar patterns when considering exits classified by the number of portfolio companies rather than the number of rounds of financing. 13% of companies sourcing funds from a community VC firm at least once have IPO exits compared with 7% of companies who never have community VC financing. Finally, 78% of all rounds with a community VC proceed to a next round of financing compared to 65% of the rounds with no community VC.

We turn to multivariate models next. While Appendix B describes the control variables

[15] A key issue is the lack of strong identifying natural experiments or instruments (although see Gompers et al. 2012 for an effort). A related issue is the long gestation period prior to observing VC outcomes. Thus, many variables are likely have dynamic effects. Finally, there is a multiplicity of counterfactuals. The structural approach of Sorensen (2007) is likely more profitable avenue for causal inference.

in detail, we briefly comment on the key controls here. First, we include a dummy variable for syndication, following the robust finding in the VC literature (e.g., Brander, Antweiler, and Amit (2002)) that syndication predicts success. We do not attempt to decompose syndication into its selection and ex-post value add components. As Sorensen (2007) points out, disentangling these effects is a complex structural exercise due to the multiplicity of counterfactuals. We follow virtually all received VC literature and take a similarly agonistic approach. We control for syndication without specifying the channels through which it acts.

We control for the stage of financing through the variable *Early Stage* and include a control for whether the portfolio firm is in the geographical clusters of California or Massachusetts. Following Chen, Gompers, Kovner, and Lerner (2010), we include related controls for whether VCs are headquartered in these agglomerates. We control for VC experience and skill (see Krishnan and Masulis (2011) for a survey). We include assets under management, centrality, *IPO Rate*, or the rate at which a VC takes its firms public, and *Experience*, the average age of the participating VCs as of the year before the financing round. We control for whether VCs are arms of financial institutions or corporate VC investors. We capture VC focus through the variables *Early Stage Focus* and *Industry Focus*, which are fractions of firms in the focus areas funded by the VC syndicate. All models include year and industry fixed effects.

5.6.1 Next Round Financing

As a starting point, we consider follow-on financing as a measure of success. Follow-on rounds of financing involve reassessment of the portfolio company. Thus, attracting follow-on funding can be viewed as one metric of success (Hochberg, Ljungqvist, and Lu (2007)). Our sample includes all rounds that are identifiably numbered and have no missing data for subsequent rounds when one exists.[16]

Table 9 reports probit estimates for the earlier of the progression from one round to another, or exit through IPO or merger within 10 years of the financing round. We find

[16]These criteria reduce the sample of first three rounds from 22,683 to 22,271 rounds. Missing rounds are spread evenly through the sample period and in both early and non-early stages.

that community VCs increase the odds of a subsequent round financing after the first and second round financings. This is unsurprising. Familiar, trusted partners likely matter more in the early stages of a venture when more effort is required to screen potential ventures. Interestingly, VC centrality has complementary effects. It is significant in later rounds but not in earlier rounds, which is in contrast to the importance of community in the first round. Perhaps thick rolodexes are more critical in later stages when it provides firms access to a broader set of resources such as personnel or strategic contacts. Familiar partners matter more in earlier rounds that call for more intensive screening.

Among the other control variables, rounds with larger VC firms have a higher probability of follow-on financing or exit. The coefficient on the early stage dummy variable is positive and statistically significant. One interpretation of this finding is that staged financing is more prevalent at the early stage firms given the greater informational issues with these firms (Cornelli and Yosha (2003)). Thus, VC firms manage early stage financing through more frequent injections of smaller amounts of capital. Syndicated deals have a higher chance of attracting future funding. Early stage focus is associated with a greater likelihood of follow-on financing or exit. Thus, firms that declare specialization in early stage ventures appear to accelerate a firm's progress to a next round of financing

5.6.2 Exit

We next consider exit as a measure of investing success. Table 10 reports the results. The baseline is a Cox proportional hazards model in which success is exit through either M&A or an IPO. We also consider a probit model in which success is exit by M&A or IPO by year 10, and a model in which exit is defined as achieving an IPO. In the IPO model, we estimate a competing hazards model that defines IPOs as successes but recognizes that IPOs are observed only conditional on not having a prior M&A.[17] We report Cox estimates in the form of an exponentiated hazards ratio where a coefficient greater than 1.0 for a variable indicates that it speeds exit. The key independent variable of interest is the *Community*

[17]We get similar results if we consider IPOs alone without accounting for the selection effect due to a (non)merger when there is an IPO.

dummy, which equals 1.0 if the round has at least one community VC, and zero otherwise.

In specification (1) of Table 10, the baseline Cox hazards model, the hazard ratio for community VC is 1.09 and is significant at the 1% level. Thus, community VC financing speeds exit. Several control variables reaffirm prior work. Among these is a control for whether a deal is syndicated or not. While it is significant – in fact, it is the economically most significant variable – community membership remains significant. Early stage ventures are slower to exit, reflecting their relative immaturity. Centrality is significant and speeds exit as in Hochberg, Ljungqvist, and Lu (2007). The geographic cluster variables are significant both for the portfolio firm and for the VC firm, consistent with Chen et al. (2010). We conduct a robustness test with regard to p-values. Instead of the statistical standard errors, we conduct placebo falsification tests. Here, we simulate communities of size and number equal to the actual numbers in the data and derive p-values based on the simulations. The inferences are robust.

The probit results in specification (2) and the competing hazards in specification (3) largely mimic the Cox results. In specifications (4) and (5), we show results classified by the round of financing. As before, community effects are pronounced in round 1, where sourcing financing from a community VC speeds exit by 10%. The screening of firms is likely more intensive in the first round when all syndicate partners confront the portfolio firm for the first round. Trusted partners sourced from communities appear to matter more in these rounds than in subsequent ones.

5.7 Other Robustness Tests

We consider the following robustness tests for the performance results. We reestimate the performance regressions with community VC financed rounds defined as portfolio companies that receive funding from *multiple* VCs from the same community in a given financing round. We define a financing round to be a community-based round if and only if there are at least 2 same-community VCs in that round. Community VC rounds are now associated with 17% faster exit, in comparison to 9% in our main specification. The results are significant at the 1% level. We also consider as community VCs only the set of VCs who belong

to communities in two (or three) successive five year periods to rule out sampling errors. While these definitions result in smaller samples of community VCs, the results are again qualitatively similar.[18]

One concern with our results is that they overstate similarity within communities because prior round syndicate members are often automatic participants in future financing rounds. Table 11 presents two sets of results to address this issue. In one test, we only consider first time deals within each 5-year rolling window to assess similarity between VCs. Panel A gives the results. None of the results are qualitatively different from those in Table 6.

In a second test, we consider syndicate relationships derived only from first round financing interactions. That is, we feed into the community detection algorithm only the first round syndication partnerships and reestimate the number and size of communities in each period. We replicate all the tables in the paper. The results in Panel B of Table 11 are qualitatively similar to those in Table 6. Community members are similar in terms of functional style and show disassortative matching on dimensions of influence.[19]

6 Conclusion

Syndication is a pervasive feature of venture capital financing. About two-thirds of the venture capital financing raised in the U.S. market is syndicated. Syndication is a robust determinant of successful VC exit. Over a period of time, a venture capital firm is likely to form syndicates several times and do so with different partners.

We study the *composition* of syndicates. Do VCs pick syndicate partners at random? Or, do they have preferred partners, and if so, is the preference assortative or disassortative? Our study provides new evidence on these questions. We find that VCs do not pick syndicate partners at random. Nor do they associate with a fixed set of partners. Rather, VCs exhibit associative preferences in which they are probabilistically more likely to syndicate with some

[18]In the Cox model, we also obtain similar results when we experiment with definitions of community VCs as VCs who belong to communities in two successive five year periods *and* belong to the same communities in two periods.

[19]Interestingly, there is dissimilarity in all three dimensions of influence: age, assets under management, and centrality.

VCs than others, which leads to clusters or spatial agglomerates that we term "communities."

We identify several communities in the VC industry using 20 years of venture financing data. We employ a flexible computational method that does not fix the number of communities, or the size of communities, and permits VCs to syndicate both within and outside their preferred clusters. About 20% of all VCs cluster into communities. We find that communities are both stable and tight-knit, suggesting that VCs have strong revealed preference for familiar syndicate partners.

Our findings highlight that familiarity can help manage the complexities of syndicated partnerships. While syndication is beneficial as it permits risk-sharing, access to diverse resources, and second opinions on risky investments, it can also pose a fresh set of problems for venture capitalists. Suspicions of ex-post hold up and free riding by partners can lead to insufficient effort and undo the benefits of syndication. Syndicating with familiar partners can mitigate these problems by reducing information asymmetry, building trust, and enhancing reciprocity between partners. Alternatively or additionally, familiar partners can enhance learning in models in which VCs learn by doing. The propensity to pick preferred syndicate partners can be interpreted as an outcome of these forces.

We also analyze the attributes of VCs *within* communities to assess the nature of partner preferences in VC syndication. We find complex but unsurprising behavioral preferences underlying partner preferences. Preferred partners are similar along dimensions of functional style such as industry or stage focus. These findings are consistent with the view that syndicate partners perform a vetting function, and that vetting is more important from functionally similar style partners. We find heterogeneity on dimensions such as VC size and influence, suggesting heterogeneity in partner preference is mainly driven by the need to extend a VC's reach. Finally, we show that community VC financed rounds are more likely to exit successfully.

The spatial locations of VC communities shed light on the nature of competition between VC syndicates. One view is that VC communities pool diverse resources to attract a wide range of portfolio firms and diverse needs for financing and post-financing support. A second view is that VC investing requires specialized skills so different VC clusters pool different

types of expertise or hard-to-acquire skills. We find evidence of the latter. VCs appear to specialize through differentiation suggesting that knowledge of local market conditions and industries constitute defensible entry barriers in the VC industry.

Our approach towards analyzing the existence and nature of partner preferences has many applications outside VCs. In finance, syndications are even more pervasive in the commercial banking area and the investment banking field, where, for instance, syndicates form for underwriting public issues. The nature of partner preferences in these areas constitute a profitable avenue for future research. Yet another area of potential concerns the theory of the firm, specifically inter-firm alliances. Robinson (2008) highlights that simultaneous, non-overlapping inter-firm collaborations are quite common even between firms that compete in some product markets.[20] An interesting question is whether these types of collaborations also exhibit preferred-partner clustering as in the VC market. Our study provides a technique for analyzing such questions and understanding the behavioral preferences that drive partnerships in these networks.

Finally, communities appear to be interesting organizational forms that lie in between formal conglomerates and firms demarcated by legal organizational boundaries. Spot contracting between legally separate entities helps avoid the inflexibility and complexity of running large conglomerates. However, it also compromises the benefits of soft information flows and relationships from an integrated conglomerate. Communities can be regarded as organizational intermediates that provide some benefits of both forms of organizations, lying somewhere in between hard-boundary conglomerates that internalize all transactions and arms-length spot contracting with outside partners.

[20] An interesting example is Apple Inc. and Samsung Corporation, who are simultaneously collaborators and competitors, or competitive collaborators, and currently engaged in litigation.

*

A. Calculating Modularity

In order to offer the reader a better sense of how modularity is computed in different settings, we provide a simple example here, and discuss the different interpretations of modularity that are possible. The calculations here are based on the measure developed in **?**. Since we used the `igraph` package in R, we will present the code that may be used with the package to compute modularity.

Consider a network of five nodes $\{A, B, C, D, E\}$, where the edge weights are as follows: $A : B = 6$, $A : C = 5$, $B : C = 2$, $C : D = 2$, and $D : E = 10$. Assume that a community detection algorithm assigns $\{A, B, C\}$ to one community and $\{D, E\}$ to another, i.e., only two communities. The adjacency matrix for this graph is

$$\{A_{ij}\} = \begin{bmatrix} 0 & 6 & 5 & 0 & 0 \\ 6 & 0 & 2 & 0 & 0 \\ 5 & 2 & 0 & 2 & 0 \\ 0 & 0 & 2 & 0 & 10 \\ 0 & 0 & 0 & 10 & 0 \end{bmatrix}$$

The Kronecker delta matrix that delineates the communities will be

$$\{\delta_{ij}\} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 1 \end{bmatrix}$$

The modularity score is

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{d_i \times d_j}{2m} \right] \cdot \delta_{ij} \tag{2}$$

where $m = \frac{1}{2} \sum_{ij} A_{ij} = \frac{1}{2} \sum_i d_i$ is the sum of edge weights in the graph, A_{ij} is the (i, j)-th entry in the adjacency matrix, i.e., the weight of the edge between nodes i and j, and $d_i = \sum_j A_{ij}$ is the degree of node i. The function δ_{ij} is Kronecker's delta and takes value

1 when the nodes i and j are from the same community, else takes value zero. The core of the formula comprises the modularity matrix $\left[A_{ij} - \frac{d_i \times d_j}{2m}\right]$ which gives a score that increases when the number of connections within a community exceeds the expected proportion of connections if they are assigned at random depending on the degree of each node. The score takes a value ranging from -1 to $+1$ as it is normalized by dividing by $2m$. When $Q > 0$ it means that the number of connections within communities exceeds that between communities. The program code that takes in the adjacency matrix and delta matrix is as follows:

```
#MODULARITY
Amodularity = function(A,delta) {
    n = length(A[1,])
    d = matrix(0,n,1)
    for (j in 1:n) { d[j] = sum(A[j,]) }
        m = 0.5*sum(d)
        Q = 0
        for (i in 1:n) {
            for (j in 1:n) {
                Q = Q + (A[i,j] - d[i]*d[j]/(2*m))*delta[i,j]
            }
        }
        Q = Q/(2*m)
    }
```

We use the R programming language to compute modularity using a canned function, and we will show that we get the same result as the formula provided in the function above. First, we enter the two matrices and then call the function shown above:

```
> A = matrix(c(0,6,5,0,0,6,0,2,0,0,5,2,0,2,0,0,0,2,0,10,0,0,0,10,0),5,5)
> delta = matrix(c(1,1,1,0,0,1,1,1,0,0,1,1,1,0,0,0,0,0,1,1,0,0,0,1,1),5,5)
> print(Amodularity(A,delta))
[1] 0.4128
```

We now repeat the same analysis using the R package. Our exposition here will also show how the walktrap algorithm is used to detect communities, and then using these communities, how modularity is computed. Our first step is to convert the adjacency matrix into a graph for use by the community detection algorithm.

```
> g = graph.adjacency(A,mode="undirected",weighted=TRUE,diag=FALSE)
```

We then pass this graph to the walktrap algorithm:

```
> wtc=walktrap.community(g,modularity=TRUE,weights=E(g)$weight)
> res=community.to.membership(g,wtc$merges,steps=3)
> print(res)
$membership
[1] 0 0 0 1 1

$csize
[1] 3 2
```

We see that the algorithm has assigned the first three nodes to one community and the next two to another (look at the membership variable above). The sizes of the communities are shown in the size variable above. We now proceed to compute the modularity

```
> print(modularity(g,res$membership,weights=E(g)$weight))
[1] 0.4128
```

This confirms the value we obtained from the calculation using our implementation of the formula.

Modularity can also be computed using a graph where edge weights are unweighted. In this case, we have the following adjacency matrix

```
> A
     [,1] [,2] [,3] [,4] [,5]
[1,]    0    1    1    0    0
```

```
[2,]    1    0    1    0    0
[3,]    1    1    0    1    0
[4,]    0    0    1    0    1
[5,]    0    0    0    1    0
```

Using our function, we get

```
> print(Amodularity(A,delta))
[1] 0.22
```

We can generate the same result using R:

```
> g = graph.adjacency(A,mode="undirected",diag=FALSE)
> wtc = walktrap.community(g)
> res=community.to.membership(g,wtc$merges,steps=3)
> print(res)
$membership
[1] 1 1 1 0 0

$csize
[1] 2 3

> print(modularity(g,res$membership))
[1] 0.22
```

A final variation on these modularity calculations is to use a Kronecker delta matrix that has diagonal elements of zero. In the paper we use the first approach presented in this Appendix.

B. Variable Definitions

Variable	Description
Age	Number of years between a VC's last investment in year t and the VC firm's founding year
AUM	Total capital under management, in $ million, of all those VC funds that invested during a 5-year rolling window
AUM_Round	natural log of one plus the average AUM, in $ million, of the participating VCs' funds that invested until the year prior to the financing round
Centrality	VC's eigenvector centrality based on syndicated rounds during a 5-year rolling window
Community	Equals 1.0 if there is at least one community VC in the financing round and zero otherwise
Company Geographical Cluster	Equals 1.0 if the portfolio company funded by the VC is in the state of California or Massachusetts and zero otherwise
Company Region HHI	Herfindahl-Hirschman index based on a VC's (or a community of VCs') share of total deals in each geographical region during each 5-year rolling window
Company Region Rank i	Geographic region with the i^{th}-highest aggregate $ amount invested by all VCs in a 5-year rolling window
Company Region Variation	Squared deviation of the proportion of a VC's (or a community of VCs') deals in a geographic region from the average of all VCs' (or all communities') proportions in the region, averaged across all VCs (or communities) and regions, during each 5-year rolling window
Corporate VC	Equals 1.0 if there is at least one venture capitalist who is the corporate VC arm of a firm
Early Stage	Equals 1.0 if the round is an early stage financing and zero otherwise
Early Stage Focus	natural log of one plus the proportion of companies that the participating VCs invested at an early stage until the year prior to the financing round
Experience	natural log of one plus the average age, in years, of the participating VCs from their founding until the year prior to the financing round[21]
FI VC	Equals 1.0 if there is at least one financial institution VC in the round
Industry Focus	natural log of one plus the proportion of companies funded by the participating VCs in the same industry as the portfolio company until the year prior to the financing round
Industry HHI	Herfindahl-Hirschman index based on a VC's (or a community of VCs') share of total deals in each industry during each 5-year rolling window

[21]Our definition modifies Lindsey(2008)'s definition on two fronts. First, we consider age based on the VC firm's founding year rather than its entry into Venture Economics. Second, we consider a VC's experience based on time periods prior to the financing round in question.

B. Variable Definitions - Contd.

Variable	Description
Industry Rank i	Industry with the i^{th}-highest aggregate $ amount invested by all VCs in a 5-year rolling window
Industry Variation	Squared deviation of the proportion of a VC's (or a community of VCs) deals in an industry from the average of all VCs' (or all communities') proportions in the industry, averaged across all VCs (or communities) and industries, during each 5-year rolling window
IPO Rate	natural log of one plus the average of each participating VC's ratio of IPOs to number of portfolio companies invested in the last three years prior to the financing round[22]
Ownership HHI	Herfindahl-Hirschman index based on the proportion of VCs in a community from each ownership type (e.g., independent private equity, corporate VC, financial institution VC arm, others)
Stage HHI	Herfindahl-Hirschman index based on a VC's (or a community of VCs') share of total deals in each stage during each 5-year rolling window
Stage Rank i	Financing stage with the i^{th}-highest aggregate $ amount invested by all VCs in a 5-year rolling window
Stage Variation	Deviation of the proportion of a VC's (or a community of VCs) deals in a stage from the average of all VCs' (or all communities') proportions in the stage, averaged across all VCs (or communities) and stages, during each 5-year rolling window
Syndicated	Equals 1.0 if the round is syndicated, zero otherwise
VC Geographical Cluster	Equals 1.0 if at least one participating VC is in the state of CA or MA
VC MSA HHI	Herfindahl-Hirschman index based on the proportion of VCs in a community from each MSA
VC Region HHI	Herfindahl-Hirschman index based on the proportion of VCs in a community from each geographic region
VC State HHI	Herfindahl-Hirschman index based on the proportion of VCs in a community from each U.S. state

[22]Krishnan and Masulis (2011)

References

Barber, B., Lyon, J. (1997). Detecting long-run abnormal stock returns: The empirical power and specification of test statistics, *Journal of Financial Economics* 43, 341-372.

Bengtsson, O., Hsu, D. (2010). How do venture capital partners match with startup founders? *Working Paper*.

Bhagwat, V. (2011). Manager networks and investment syndication: Evidence from venture capital. *Working Paper*.

Bottazzi, L., Da Rin, M., Hellmann, T. (2011). The importance of trust for investment: Evidence from venture capital. *Working Paper*.

Brander, J. A., Amit, R., Antweiler, W. (2002). Venture-capital syndication: Improved venture selection vs. the value-added hypothesis, *Journal of Economics and Management Strategy*, v11, 423-452.

Brown, S., Goetzmann, W. (1997). Mutual fund styles. *Journal of Financial Economics* 43, 373-399.

Brown, S., Warner, J. (1985) Using daily stock returns: The case of event studies. *Journal of Financial Economics* 14, 3-31.

Burdick, D., Hernandez, M., Ho, H., Koutrika, G., Krishnamurthy, R., Popa, L., Stanoi, I.R., Vaithyanathan, S., Das, S.R. (2011). Extracting, linking and integrating data from public sources: A financial case study, *IEEE Data Engineering Bulletin*, 34(3), 60-67.

Bygrave, W.D., Timmons, J.A. (1992). Venture capital at the crossroads. Harvard Business Press.

Cestone, Giacinta., Lerner, Josh, White, Lucy (2006). The design of syndicates in venture capital, *Harvard Business School Working Paper*.

Chen, Henry, Gompers, Paul, Kovner, Anna, Lerner, Josh (2010). Buy local? The geography of successful venture capital expansion, *Journal of Urban Economics* 67(1).

Chidambaran, N. K., Kedia, S., Prabhala, N.R. (2010). CEO-Director connections and fraud, *University of Maryland Working Paper*.

Chung, Seungwha, Singh, H., Lee, K. (2000). Complementarity, Status-Seeking, and Social Capital As Drivers of Alliance Formation, *Strategic Management Journal* 21(1), 1-22.

Cohen, Lauren, Frazzini, Andrea, Malloy, Christopher (2010). Sell-Side school ties, *Journal of Finance* 65, 1409-1437.

Cohen, Lauren, Frazzini, Andrea, Malloy, Christopher (2012). Hiring cheerleaders: Board Appointments of Independent Directors, forthcoming, *Management Science*.

Cornelli, F., Yosha, O. (2003). Stage financing and the role of convertible securities, *Review of Economic Studies* 70, 1-32.

Currarini, Sergio., Jackson, Matthew., Pin, Paolo. (2012).An Economic Model of Friendship: Homophily, Minorities, and Segregation, *Econometrica* 77, 1003-1045.

Da Rin, Marco, Hellmann, Thomas, Puri, Manju (2012). A survey of venture capital research, *Duke University Working Paper*.

Du, Qianqian (2011). Birds of a feather or celebrating differences? The formation and impact of venture capital syndication, *University of British Columbia Working Paper*.

Engelberg, Joseph., Gao, Pengjie, Parsons, Christopher (2010). The value of a rolodex: CEO pay and personal networks, *Review of Financial Studies*, forthcoming.

Fortunato, S. (2009). Community detection in graphs, *arXiv*:0906.0612v1 [physics.soc-ph].

Gertler, M.S. (1995). Being there: proximity, organization and culture in the development and adoption of advanced manufacturing technologies, *Economic Geography* 7(1), 1-26.

Goldfarb, Brent, Kirsch, David, Miller, David, (2007). Was there too little entry in the dot com era?, *Journal of Financial Economics* 86(1), 100-144.

Gompers, P., Lerner, J. (2001). The venture capital revolution, *Journal of Economic Perspectives* 15(2), 45-62.

Gompers, P., Mukharlyamov, V., Xuan, Y. (2012). The cost of friendship, *Harvrad Business School Working Paper*.

Gorman, M., Sahlman, W. (1989). What do venture capitalists do? *Journal of Business Venturing* 4, 231-248.

Granovetter, M. (1973). The strength of weak ties. *American Journal of Sociology* 78(6), 1360-1380.

Granovetter, M. (1985). Economic action and social structure: The problem of embeddedness, *American Journal of Sociology* 91(3), 481-510.

Grossman, S., Hart, O. (1986). The costs and benefits of ownership: A theory of vertical and lateral integration, *Journal of Political Economy* 94(4), 691-719.

Guiso, L., Sapienza, P., Zingales, L. (2004). The role of social capital in financial development, *American Economic Review* 94, 526-556.

Gulati, R. (1995). Does familiarity breed trust? The implications of repeated ties for contractual choice in alliances, *The Academy of Management Journal* 38(1).

Harrison, D., Klein, K. (2007). What's the difference? Diversity constructs as separation, variety, or disparity in organization, *Academy of Management Review* 32(4), 1199-1228.

Hart, O., Moore, J. (1990). Property rights and the nature of the firm, *Journal of Political Economy* 98(6), 1119-1158.

Hegde, D., Tumlinson, J. (2011). Can birds of a feather fly together? Evidence for the economic payoffs of ethnic homophily, *Working Paper*.

Hellmann, T. J., Puri, M. (2002). Venture capital and the professionalization of start-up firms: Empirical evidence, *Journal of Finance* 57, 169-197.

Hoberg, G., Phillips, G. (2010). Product Market Synergies and Competition in Mergers and Acquisitions: A Text-Based Analysis, *Review of Financial Studies* 23(10), 3773–3811.

Hochberg, Y., Ljungqvist, A., Lu, Y. (2007). Whom you know matters: Venture capital networks and investment performance, *Journal of Finance* 62(1), 251-301.

Hochberg, Y., Lindsey, L., Westerfield, M. (2012). Partner selection in co-investment networks: Evidence from venture capital. *Northwestern University Working Paper*.

Hsu, David. (2004). What Do Entrepreneurs Pay for Venture Capital Affiliation?, *Journal of Finance* 59, 1805-1844.

Hwang, B., Kim, S. (2009). It pays to have friends, *Journal of Financial Economics* 93, 138-158.

Kaplan, S. N., Sensoy, B., Stromberg, P. (2002). How well do venture capital databases reflect actual investments?, *Working paper*, University of Chicago.

Kaplan, S. N., Stromberg, P. (2003). Financial contracting theory meets the real world: Evidence from venture capital contracts, *Review of Economic Studies* 70, 281-316.

Kaplan, S. N., Stromberg, P. (2004). Characteristics, contracts and actions: Evidence from venture capital analyses, *Journal of Finance* 59, 2177-2210.

Krishnan, C. N. V., Masulis, R. W. (2011). Venture capital reputation, in Douglas J. Cummings, ed., *Handbook on Entrepreneurial Finance, Venture Capital and Private Equity*, Oxford University Press.

Lerner, J., Hardymon, F., Leamon, A. (2007). Venture capitalist and private equity: A casebook, John Wiley and Sons.

Lindsey, L. A. (2008). Blurring boundaries: The role of venture capital in strategic alliances, *Journal of Finance* 63(3), 1137-1168.

Lucas, R., 1978. On the size distribution of business firms. *Bell Journal of Economics* 9, 508–523.

Maats, F., Metrick, A., Hinkes, B., Yasuda, A., Vershovski, S. (2008). On the Completeness and Interchangeability of Venture Capital Databases, *UC Davis Working paper*

Maksimovic, V., Phillips, G. 2002. Do conglomerate firms allocate resources inefficiently across industries? *Journal of Finance* 57, 721–776.

McPherson, M., Smith-Lovin, L., Cook, J. (2001). Birds of a feather: Homophily in social networks, *Annual Review of Sociology* 27, 415-444.

Neher, D. V. (1999). Staged financing: An agency perspective, *Review of Economic Studies* 66, 255-274.

Newman, M. (2001). Scientific collaboration networks: II. Shortest paths, weighted networks, and centrality, *Physical Review E* 64, 016132.

Newman, M. (2010). Networks: An introduction, Oxford University Press.

Pons, Pascal, Latapy, Matthieu, (2005). Computing communities in large networks using random walks, *Journal of Graph Algorithms and Applications* 10, 191–218.

Porter, M.E. (2000). Location, competition and economic development: Local clusters in a global economy, *Economic Development Quarterly* 14(1), 15-34.

Porter, Mason, Mucha, Peter, Newman, Mark, Friend, A. J. (2007). Community structure in the United States House of Representatives, *Physica A: Statistical Mechanics and its Applications* 386(1), 413–438.

Roberts, Michael, Whited, Toni. (2011). Endogeneity in Empirical Corporate Finance, *Handbook of the Economics of Finance Volume 2, Elsevier*, forthcoming.

Robinson, D. (2008). Strategic alliances and the boundaries of the firm, *Review of Financial Studies* 21(2), 649-681.

Robinson, D., Sensoy, B. (2011). Private equity in the 21st century: cash flows, performance, and contract terms from 1984-2010, *Working Paper*, Ohio State University.

Robinson, D., Stuart, T. (2007). Financial contracting in biotech strategic alliances, *Journal of Law and Economics* 50(3), 559-596.

Sorensen, Morten (2007). How smart is smart money? A Two-sided matching model of venture capital, *Journal of Finance* 62, 2725-2762.

Sorensen, Morten (2008). Learning by investing: evidence from venture capital, *Columbia University Working Paper*.

Sorenson, O., Stuart, T. (2001). Syndication networks and spatial distribution of venture capital investment, *American Journal of Sociology* 106, 1546-1588.

Stanfield, Jared (2013). How does syndication influence leveraged buyout success?, *University of New South Wales Working Paper*.

Tian, Xuan (2011). The causes and consequences of venture capital stage financing, *Journal of Financial Economics* 101(1), 132–159.

Uzzi, B. (1997). Social structure and competition in interfirm networks: The paradox of embeddedness. *Administrative Science Quarterly* 42, 35-67.

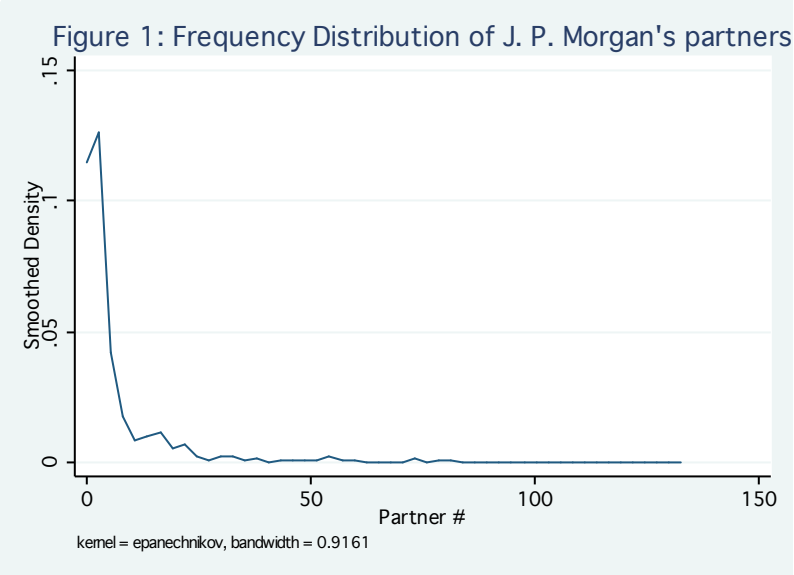


Figure 1: Frequency distribution of J.P. Morgan's partners.

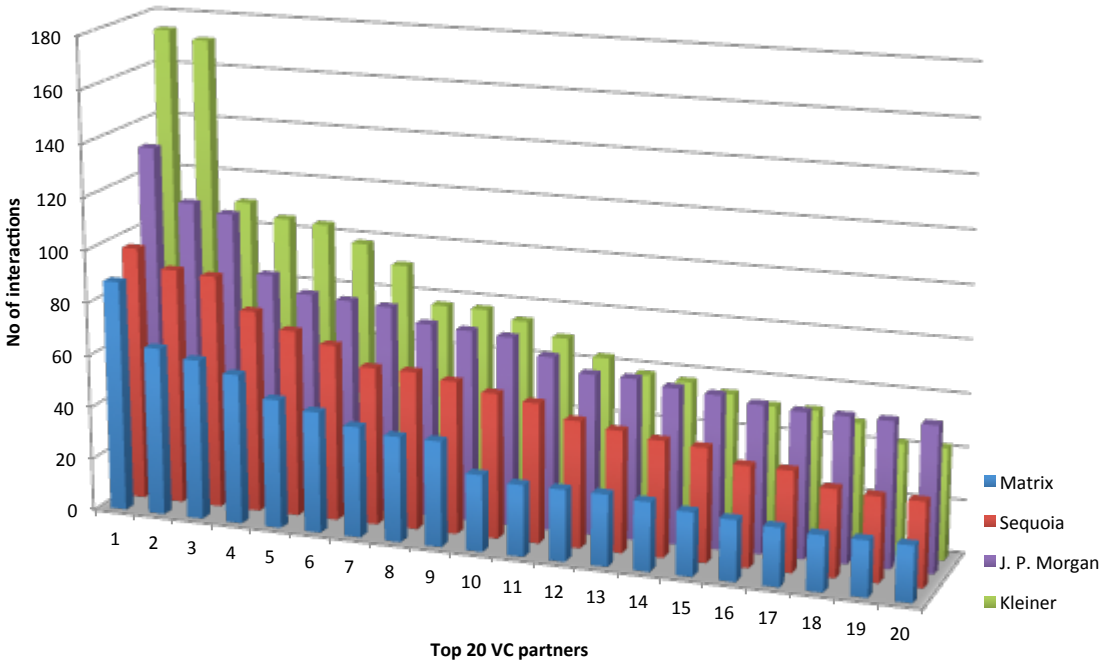


Figure 2: Distribution of the number of interactions of four top firms with their top 20 collaborators.

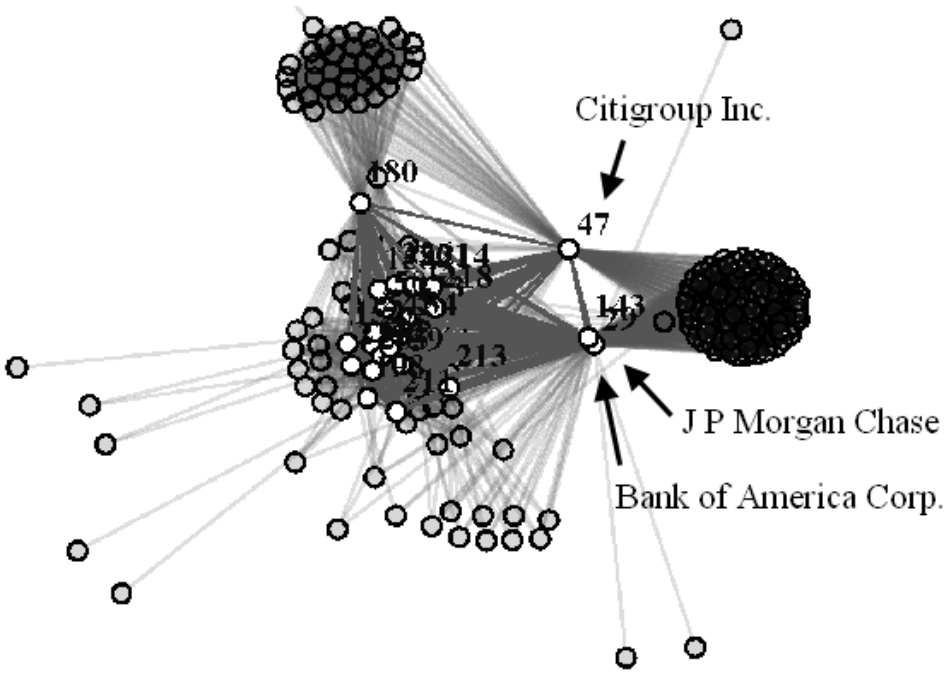


Figure 3: Communities and centrality in bank co-lending networks.

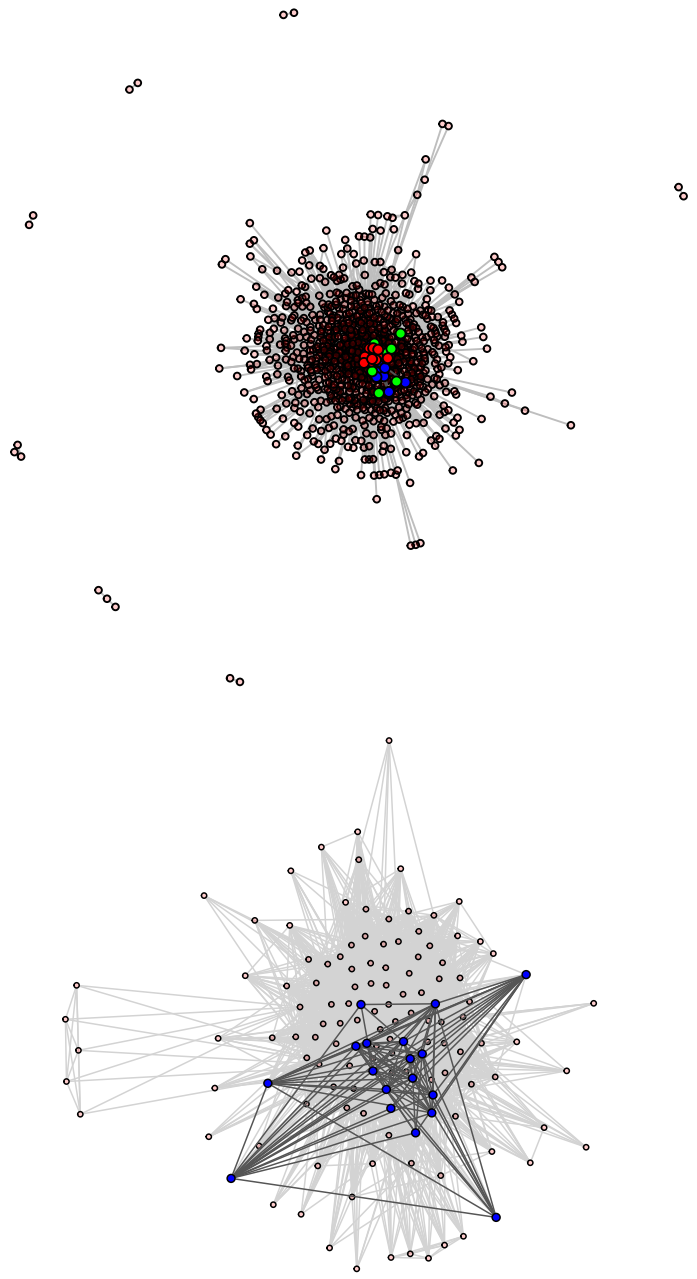


Figure 4: Network graph for connected VCs (1980–84). The upper plot shows the network of all VCs in communities (1180 in all), and blue, green, and red nodes in the center of the network are the VCs in the top three largest communities, respectively. The lower plot shows the network comprised only of the 134 VCs who are members of the 18 communities that have at least five VCs. The darker nodes in the lower plot show the VCs in the largest community.

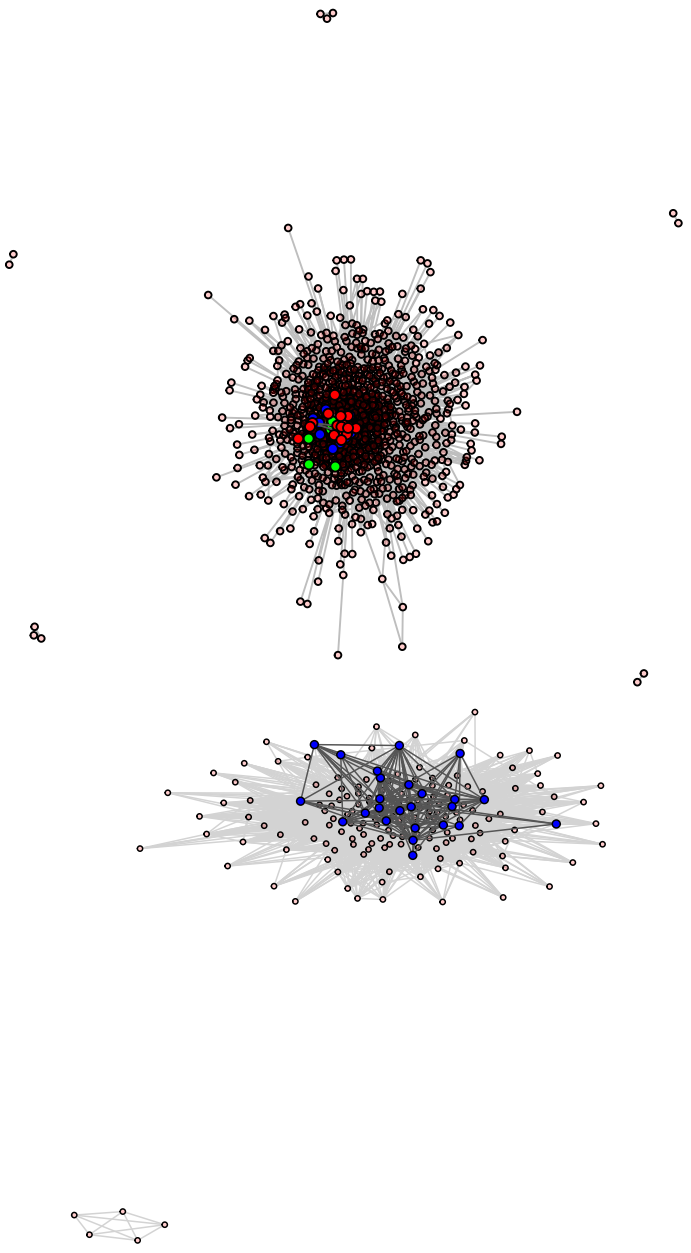


Figure 5: Network graph for connected VCs (1985–89). The upper plot shows the network of all VCs in communities (1295 in all), and blue, green, and red nodes in the center of the network are the VCs in the top three largest communities, respectively. The lower plot shows the network comprised only of the 180 VCs who are members of the 18 communities that have at least five VCs. The darker nodes in the lower plot show the VCs in the largest community. Note the single satellite community at the bottom of the lower plot. Such a community has low centrality.

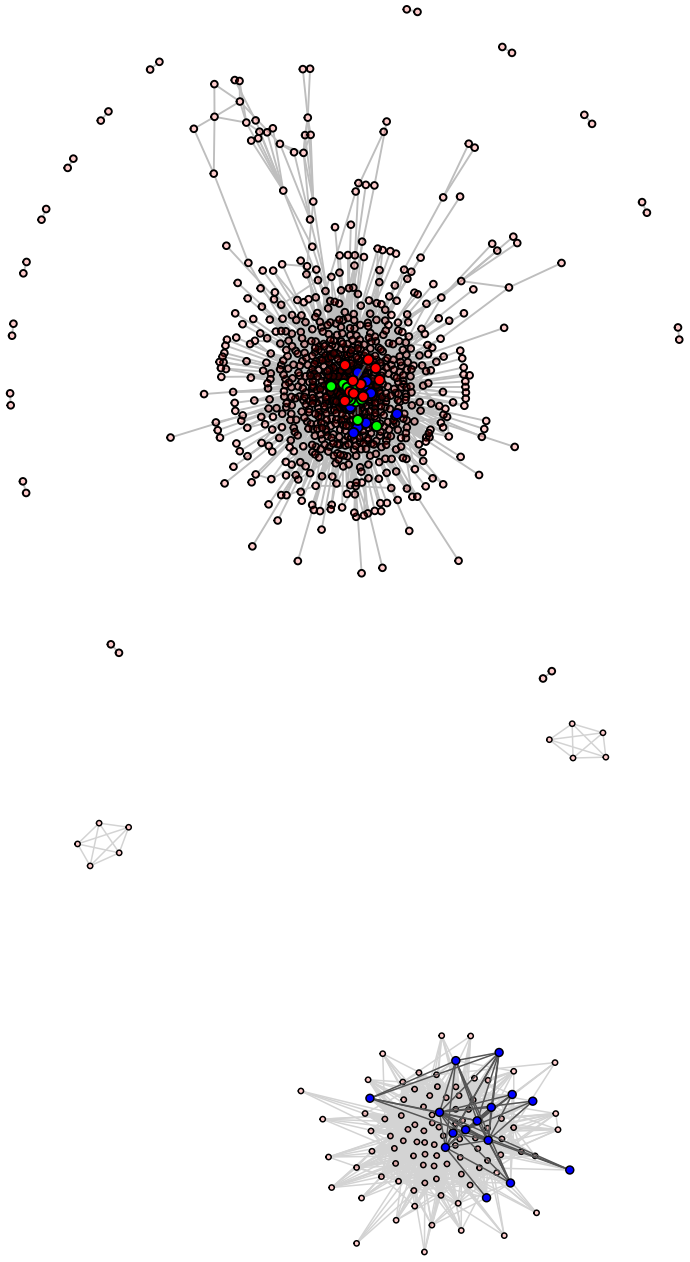


Figure 6: Network graph for connected VCs (1990–94). The upper plot shows the network of all VCs in communities (953 in all), and blue, green, and red nodes in the center of the network are the VCs in the top three largest communities, respectively. The lower plot shows the network comprised only of the 114 VCs who are members of the 14 communities that have at least five VCs. The darker nodes in the lower plot show the VCs in the largest community. Note the two satellite communities above the main one in the lower plot. Such communities have low centraity.

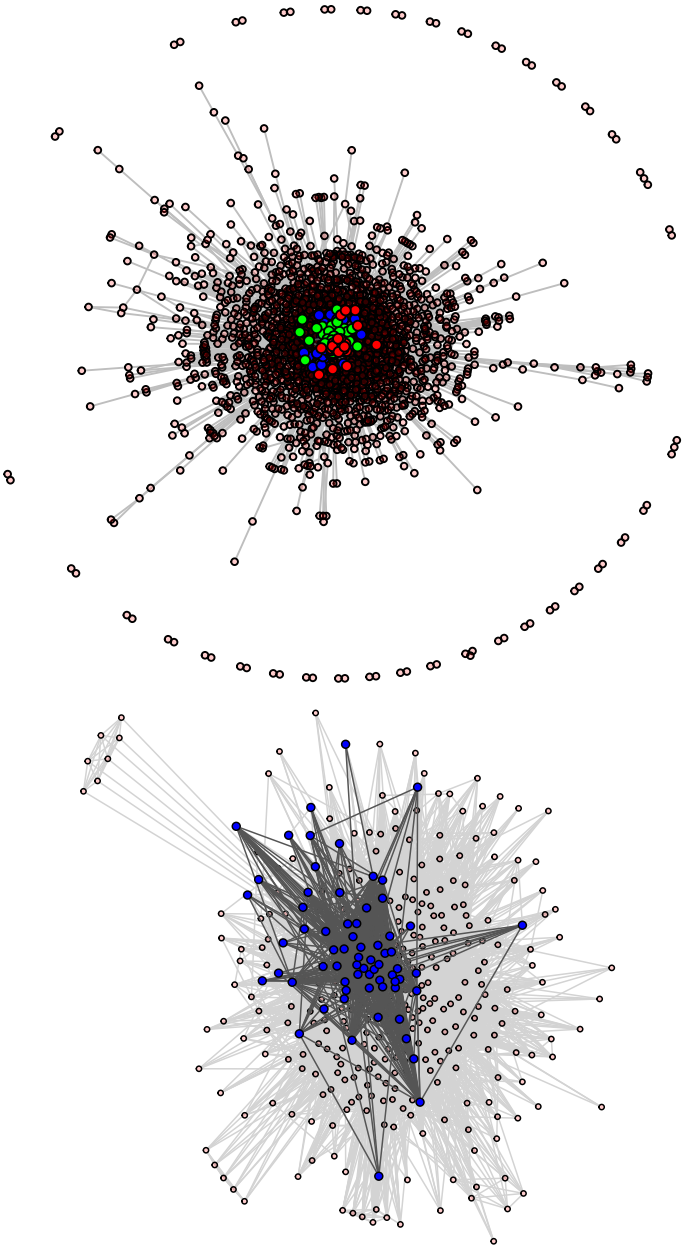


Figure 7: Network graph for connected VCs (1995–99). The left plot shows the network of all VCs in communities (2772 in all), and blue, green, and red nodes in the center of the network are the VCs in the top three largest communities, respectively. The right plot shows the network comprised only of the 379 VCs who are members of the 35 communities that have at least five VCs. The darker nodes in the right plot show the VCs in the largest community.

Table 1: Venture Capitalists in our sample. This table provides descriptive statistics of the 1,962 unique U.S.-based VCs in our database over the entire 20-year period, from 1980 to 1999. Data are from Venture Economics and exclude non-US investments, angel investors, and VC firms focusing on buyouts. We report the number of rounds of financing and the count of portfolio companies a VC invests in. Investment per round is the amount a VC invests in a round. % Deals Syndicated is the number of a VC's syndicated rounds as a percentage of all rounds that a VC invested in. % Early Stage Deals is the number of a VC's investment rounds classified by Venture Economics as early stage as of the round financing date, as a percentage of all Venture Economics deals for the VC between 1980 and 1999. AUM is the sum of the capital under management of a VC in all funds that invested during 1980-1999. Total investment is the sum of a VC's investments over this time period. Age is defined as the difference in the year of the VC's last investment in the period 1980 to 1999 and the VC firm's founding date. # VC firms per MSA is the total number of unique VCs headquartered a metropolitan statistical area (MSA). CA/MA VC is the fraction of all VCs that are headquartered in either California or Massachusetts.

Variables:	Mean	Median	# Observations
# Rounds	47.98	9.00	1,962
# Companies	21.64	7.00	1,962
Investment per round ($ mm)	1.95	1.06	1,945
% Deals Syndicated	73.62	80.90	1,962
% Early Stage Deals	35.95	33.33	1,962
AUM ($ mm)	128.01	17.50	1,552
Total Investment ($ mm)	59.51	11.05	1,945
Age	9.59	6.00	1,950
# VC firms per MSA	14.24	3.00	127
CA/MA VC	0.35	0.00	1,962

Table 2: Stability of community status. The table provides data on the number of VCs who belong to community clusters in each 5-year window and the fraction of these that remain in a community after 1, 3, and 5 years from the initial window.

Window	# Community VCs	After 1 year	After 3 years	After 5 years
1980-1984	134	0.90	0.85	0.77
1981-1985	153	0.96	0.90	0.80
1982-1986	180	0.93	0.80	0.72
1983-1987	177	0.96	0.87	0.77
1984-1988	205	0.87	0.78	0.67
1985-1989	180	0.92	0.83	0.71
1986-1990	169	0.88	0.76	0.69
1987-1991	125	0.88	0.79	0.77
1988-1992	130	0.93	0.78	0.75
1989-1993	111	0.86	0.77	0.71
1990-1994	114	0.89	0.80	0.77
1991-1995	112	0.82	0.80	
1992-1996	146	0.93	0.89	
1993-1997	173	0.90		
1994-1998	246	0.94		
1995-1999	379			

Table 3: Stability of community composition. We identify community clusters in a 5-year window and examine whether the communities in the next five year window are similar to the ones in the previous period. We use the Jaccard similarity index which measures the similarity between every pair of communities in the adjacent period, and average it across all non-empty intersections. The Jaccard index is defined as the ratio of the size of the intersection set to the size of the union set. To benchmark the index, we generate a similar index for simulated communities generated by matching same community sizes and number of communities in each 5-year rolling window as in our sample. The last column shows the p-values testing the equality of the composite measure for the community and simulated community. *** denotes significance at the 1% level.

Window 1	Window 2	Community	Bootstrapped Community	p-value
1980-1984	1981-1985	0.188	0.064	0.01***
1981-1985	1982-1986	0.175	0.060	0.01***
1982-1986	1983-1987	0.182	0.056	0.01***
1983-1987	1984-1988	0.217	0.058	0.01***
1984-1988	1985-1989	0.141	0.055	0.01***
1985-1989	1986-1990	0.177	0.052	0.01***
1986-1990	1987-1991	0.155	0.052	0.01***
1987-1991	1988-1992	0.155	0.050	0.01***
1988-1992	1989-1993	0.252	0.055	0.01***
1989-1993	1990-1994	0.123	0.062	0.01***
1990-1994	1991-1995	0.246	0.065	0.01***
1991-1995	1992-1996	0.143	0.055	0.01***
1992-1996	1993-1997	0.128	0.042	0.01***
1993-1997	1994-1998	0.135	0.041	0.01***
1994-1998	1995-1999	0.109	0.042	0.01***

Table 4: Descriptive statistics for 33,924 rounds in 13,541 unique portfolio companies from 1985-1999. A round is a community round if at least one VC firm participating in it comes from a VC community. Communities are detected using a walk trap algorithm applied to syndicated deals over five year windows rolled forward one year at a time. The sample comprises VC deals obtained from Venture Economics but excludes non-US investments, angel investors and VC firms focusing on buyouts. Industry classifications are as per Venture Economics. Exit data are obtained by matching with Thomson Financial IPO and M&A databases.

Variable	Total	Community Round	Not Community Round
	Panel A: Counts By Round		
# Deals	33,924	15,220	18,704
—Round 1	11,018	3,581	7,437
—Round 2	6,881	3,015	3,866
—Round 3	4,784	2,410	2,374
Syndicated	14,897	10,056	4,841
Early stage	12,118	5,472	6,646
Geographical Cluster	16,270	9,607	6,663
Rounds with			
—Geographical Cluster VC	19,678	12,140	7,538
—Corporate VC	3,372	1,923	1,449
—FI VC	7,586	4,415	3,171
	Panel B: Percentage By Venture Economics Industry		
—Biotech	6.8	7.3	6.3
—Commu&Media	12.1	13.3	11.1
—Hardware	7.3	9.0	6.0
—Software	19.8	22.7	17.5
—Semiconductor, Electricals	7.0	7.9	6.3
—Consumer Products	7.8	5.3	9.9
—Industrial, Energy	5.9	3.4	8.0
—Internet	11.0	11.9	10.3
—Medical	13.7	15.0	12.7
—Others	8.5	4.4	11.9
	Panel C: Round Statistics		
Proceeds ($ million)	4 (1)	5 (2)	3 (1)
# VCs	2.08 (1)	2.89 (2)	1.42 (1)
—in syndicated rounds	3.46 (3)	3.85 (3)	2.64 (2)
—in early stage rounds	1.93 (1)	2.53 (2)	1.43 (1)
—in round 1	1.54 (1)	2.03 (2)	1.31 (1)
—in round 2	2.00 (1)	2.70 (2)	1.45 (1)
—in round 3	2.38 (2)	3.23 (3)	1.52 (1)
	PANEL D: Exit		
Rounds with			
—IPO exits	3,828	2,071	1,757
—M&A exits	8,794	4,363	4,431
—Follow-on funding	23,972	11,903	12,069

Table 5: Characteristics of Same-Community VCs. The table compares key community characteristics with those of simulated communities generated by matching community sizes and number of communities in each 5-year rolling window. For each community (and simulated community), we generate the mean of the characteristic, and present the average value across communities. *Age* uses the number of years between a VC's last investment in a 5-year window and the founding year of the VC firm. *Assets under management (AUM)*, in $ million, uses the sum of all VC funds that invested during a 5-year period. *Centrality* is based on each VC's eigenvector centrality determined for each 5-year rolling window. For the remaining attributes, we calculate the Herfindahl-Hirschman Index (HHI) as the sum of squared share in each subcategory of the attribute. *Industry HHI* is the Herfindahl index based on the % of a community VC's deals in each industry, while *Stage HHI* is the Herfindahl index based on the % of deals in each stage of investment. *Company Region HHI* is the Herfindahl index based on the % of deals in each geographic region. In unreported tests, we see similar results when we use HHI based on amount invested. The industry, stage and geographic region classifications are those provided by Venture Economics. The last column shows the p-values testing the equality of the means of the community and bootstrapped community characteristics. ***, **, and * denote 1%, 5% and 10% significance, respectively.

	Community	Simulated Community	p-value
Age	9.18	8.25	0.01***
AUM	130.40	70.72	0.01***
Centrality	0.08	0.03	0.01***
Industry HHI	0.28	0.48	0.01***
Stage HHI	0.33	0.52	0.01***
Company Region HHI	0.42	0.58	0.01***

Table 6: Similarity of Within-Community VCs. The table presents variation in key attributes (in Panels A-B) and mean geographic location HHI (in Panel C) and ownership HHI (in Panel D) of VCs within communities, and compares these to those of simulated communities generated by matching community sizes and number of communities in each 5-year rolling window. We calculate the Herfindahl-Hirschman index (HHI) as the sum of squared deviation of each subcategory of the attribute. *Age* uses the number of years between a VC's last investment in a 5-year rolling window and the founding year of the VC firm. *Assets under management (AUM)*, in $ million, uses the sum of all VC funds that invested during each 5-year rolling window. *Centrality* is based on each VC's eigenvector centrality determined for each 5-year rolling window. *Industry*, *Stage* and *Company Region* are based on % of a VC's deals in each of the 10 industries, each of the 5 stages, and each of the 14 U.S. geographic regions, respectively, as classified by Venture Economics. In Panels A and B, variations in Reach attributes and attribute HHI, respectively, are the standard deviation of each attribute of a community's VC, averaged across all communities. Variation in each attribute in Panel B measures the mean (across all communities) of the sum of squared deviation in each subcategory (e.g., Industry j) of each attribute (e.g., Industry) averaged across all subcategories and all within-community VCs. Panel C uses alternative geographic location variables, from the most granular (MSA) to the least granular (Region), and calculates the geographic HHI of a community's VCs, averaged across all communities. Panel D calculates the ownership HHI of a community's VCs, averaged across all communities. The last column shows the p-values testing the equality of the means of the community and simulated community characteristics. ***, **, and * denote 1%, 5% and 10% significance, respectively, from the test.

	Community	Simulated Community	p-value
	Panel A: Variation in Reach Attributes		
Age	6.86	7.37	0.01***
AUM	142.74	99.52	0.01***
Centrality	0.08	0.05	0.01***
	Panel B: Variation in Functional Styles		
Industry HHI	0.22	0.31	0.01***
Stage HHI	0.21	0.28	0.01***
Company Region HHI	0.21	0.31	0.01***
Industry Variation	0.96	3.20	0.01***
Stage Variation	0.70	2.28	0.01***
Company Region Variation	0.89	3.65	0.01***
	Panel C: Mean of Community Geographic HHI		
VC MSA HHI	0.35	0.20	0.01***
VC State HHI	0.43	0.24	0.01***
VC Region HHI	0.41	0.25	0.01***
	Panel D: Mean of Community Ownership HHI		
VC Ownership HHI	0.55	0.45	0.01***

Table 7: Functional Expertise Similarity of Within-Community VCs. We present the mean (across all communities) of the sum of squared deviation of VC's share of deal in some subcategories (based on total $ amount invested in a 5-year rolling window in each of the top 4 industries, top 2 stages, and top 4 company regions, with the remainder share of investment comprising the last subcategory in each). We compare these values to those of simulated communities generated by matching community sizes and number of communities in each 5-year rolling window. The last column shows the p-values testing the equality of the means of the community and simulated community characteristics. ***, **, and * denote 1%, 5% and 10% significance, respectively, from the test.

	Community	Simulated Community	p-value
Industry Rank:			
1	0.16	0.23	0.01***
2	0.13	0.21	0.01***
3	0.12	0.18	0.01***
4	0.11	0.17	0.01***
5 = Others	0.18	0.33	0.01***
Stage Rank:			
1	0.17	0.22	0.01***
2	0.18	0.24	0.01***
3 = Others	0.16	0.28	0.01***
Company Region Rank:			
1	0.20	0.31	0.01***
2	0.10	0.22	0.01***
3	0.11	0.17	0.01***
4	0.07	0.16	0.01***
5 = Others	0.16	0.36	0.01***

Table 8: Similarity Across Communities. The table presents across community variation in (average) key VC attributes (in Panel A), in geographic location HHI (in Panel B) and in ownership HHI (in Panel C) of VCs within communities, and compares these to those of simulated communities generated by matching community sizes and number of communities in each 5-year rolling window. We calculate the Herfindahl-Hirschman index (HHI) as the sum of squared deviation of each subcategory of the attribute. *Industry*, *Stage* and *Company Region* are based on % of a VC's deals in each of the 10 industries, each of the 5 stages, and each of the 14 U.S. geographic regions, respectively, as classificed by Venture Economics. Panel A shows the mean values (across 5-year rolling windows) of standard deviation of the within-community HHI from the across-community average in each 5-year rolling window. In addition in Panel A, for each 5-year rolling window, we calculate the squared deviation of the within-community averages from the across-community averages in each subcategory (e.g., Industry j) of each attribute (e.g., Industry), summed across all subcategories. The table presents mean of these values across all 5-year windows. Panel B uses alternative geographic location variables, from the most granular (MSA) to the least granular (Region), and calculates the standard deviation of geographic HHI of communities in each 5-year window, averaged across all such windows. Panel C calculates the standard deviation of ownership HHI of communities in each 5-year window, averaged across all such windows. The last column shows the p-values testing the equality of the means of the community and simulated community characteristics. ***, **, and * denote 1%, 5% and 10% significance, respectively, from the test.

	Community	Simulated Community	p-value
	Panel A: Variation in Functional Styles		
Industry HHI	0.14	0.04	0.01***
Stage HHI	0.11	0.05	0.01***
Company Region HHI	0.16	0.07	0.01***
Industry Variation	1.30	0.60	0.01***
Stage Variation	0.63	0.39	0.01***
Company Region Variation	1.40	1.01	0.01***
	Panel B: Variation of Community Geographic HHI		
VC MSA HHI	0.19	0.08	0.01***
VC State HHI	0.20	0.09	0.01***
VC Region HHI	0.19	0.09	0.01***
	Panel C: Variation of Community Ownership HHI		
VC Ownership HHI	0.19	0.14	0.01***

Table 9: Success through next round financing or exit. The table reports the estimates of a probit model in which the dependent variable is 1.0 if there is a successful exit (IPO or merger) or a follow-on financing round within 10 years of the investment round and 0 otherwise. See Appendix B for a description of the independent variables. All specifications include year and industry fixed effects, which are not reported for brevity. The sample comprises VC deals obtained from Venture Economics but excludes non-US investments, angel investors and VC firms focusing on buyouts. *t*-statistics based on robust standard errors are in parentheses. All specifications are overall significant at the 1% level. ***, **, and * denote significance at the 1%, 5% and 10% levels, respectively.

	Round1	Round2	Round3
	(1)	(2)	(3)
Community	0.093**	0.192***	0.033
	(2.12)	(2.79)	(0.40)
Early Stage	0.299***	0.280***	0.271***
	(9.09)	(5.03)	(3.62)
Company Geographical Cluster	0.090**	0.039	0.142**
	(2.56)	(0.67)	(2.09)
AUM_Round	0.179***	0.073***	0.106***
	(13.14)	(2.84)	(2.93)
Corporate VC	-0.066	0.026	0.131
	(-0.98)	(0.31)	(1.40)
FI VC	-0.124***	-0.081	0.019
	(-3.19)	(-1.31)	(0.25)
Syndicated	0.515***	0.558***	0.589***
	(13.87)	(9.34)	(7.71)
IPO Rate	-0.267***	-0.556***	-0.194
	(-3.57)	(-3.81)	(-0.94)
Centrality	-0.068***	0.021	0.125**
	(-3.11)	(0.61)	(2.54)
VC Geographical Cluster	0.066*	0.029	-0.116
	(1.84)	(0.46)	(-1.44)
Experience	-0.102***	-0.088**	-0.125***
	(-5.61)	(-2.48)	(-2.62)
Early Stage Focus	0.320***	0.734***	0.705**
	(2.92)	(3.23)	(2.36)
Industry Focus	0.082	0.086	0.139
	(0.72)	(0.39)	(0.48)
# Observations	9,328	4,262	3,105

Table 10: Time to exit and probability of exit. Specification (1) reports the estimates of a Cox proportional hazards model. The dependent variable is the number of days from financing to the earlier of exit (IPO or merger) or April 30, 2010. Specification (2) reports the estimates of a probit model in which the dependent variable is 1.0 if there is an exit (IPO or merger) within 10 years of the investment round and 0 otherwise. Specifications (3)-(5) report estimates of a competing hazards model where the event of interest is exit only through an IPO (Specification (3)), IPO or follow on financing after round 1 (Specification (4)) or after round 2 (Specification (5)). A merger is the competing risk in the competing hazards models. See Appendix B for a description of the independent variables. The sample comprises VC deals obtained from Venture Economics but excludes non-US investments, angel investors and VC firms focusing on buyouts. All specifications include year and industry fixed effects, which are not reported for brevity. Both the specifications are overall significant at 1%. t-statistics based on robust standard errors are in parentheses. ***, **, and * denote significance at the 1%, 5% and 10% levels, respectively.

	Cox	Probit	Competing Hazards		
			IPO	Round 1	Round 2
	(1)	(2)	(3)	(4)	(5)
Community	1.089***	0.043*	1.116**	1.095**	0.950
	(2.89)	(1.91)	(2.14)	(2.11)	(-0.84)
Early Stage	0.911***	-0.037**	0.849***	1.425***	1.375***
	(-4.00)	(-2.09)	(-3.95)	(9.69)	(6.62)
Company Geographical Cluster	1.057**	0.038**	1.060	1.078**	0.959
	(2.30)	(2.05)	(1.38)	(2.05)	(-0.83)
AUM_Round	1.088***	0.057***	1.130***	1.151***	1.048*
	(6.36)	(6.14)	(4.88)	(8.94)	(1.92)
Corporate VC	1.320***	0.202***	1.503***	0.835***	0.978
	(8.23)	(7.53)	(7.46)	(-2.66)	(-0.33)
FI VC	1.083***	0.056***	1.191***	0.897***	0.900*
	(3.01)	(2.77)	(3.90)	(-2.58)	(-1.94)
Syndicated	1.318***	0.211***	1.311***	1.386***	1.305***
	(10.52)	(10.80)	(5.77)	(9.13)	(4.57)
IPO Rate	1.084	0.063	1.145	0.692***	0.648**
	(1.26)	(1.25)	(1.23)	(-3.86)	(-2.41)
Centrality	0.998	0.006	0.983	0.943***	1.032
	(-0.19)	(0.54)	(-0.75)	(-2.77)	(1.15)
VC Geographical Cluster	1.039	0.026	1.075	1.011	1.000
	(1.38)	(1.24)	(1.44)	(0.29)	(0.01)
Experience	0.958***	-0.035***	1.002	0.919***	0.946*
	(-2.61)	(-2.91)	(0.07)	(-4.31)	(-1.73)
Early Stage Focus	1.043	0.008	0.546***	1.894***	1.850***
	(0.42)	(0.10)	(-3.20)	(5.19)	(2.90)
Industry Focus	1.090	0.062	1.542**	1.155	1.040
	(0.88)	(0.85)	(2.47)	(1.21)	(0.20)
# Observations	23,977	24,864	23,977	9,037	4,108

Table 11: Robustness of Within-Communty Similarity. This table provides 2 robustness tests of similarity among VCs within communities. Panel A considers a community VC's first investment in each portfolio company for determining % of deals in each subcategory of attributes, and uses it to determine within-comunity HHI variation as well as variation between subcategories. Panel B uses an alternative basis for communities, namely the first round of syndications rather than all rounds used in our analysis so far. Using these alternative communities, we determine within-community HHI variation and variation between subcategories. Given the alternative community, we additionally present the standard deviation of the reach variables only in Panel B. We compare these values to those of simulated communities generated by matching community sizes and number of communities in each 5-year rolling window. The last column shows the p-values testing the equality of the means of the community and simulated community characteristics. ***, **, and * denote 1%, 5% and 10% significance, respectively, from the test.

	Community	Simulated Community	p-value
	Panel A: Only First Time Deals in Each 5-Year Window		
Industry HHI	0.20	0.31	0.01***
Stage HHI	0.18	0.27	0.01***
Company Region HHI	0.21	0.31	0.01***
Industry Variation	0.96	3.20	0.01***
Stage Variation	0.86	2.48	0.01***
Company Region Variation	0.89	3.64	0.01***
	Panel B: Community detected based only on First Round Syndicates		
Age	7.58	7.22	0.10*
AUM	154.80	96.54	0.01***
Centrality	0.08	0.04	0.01***
Industry HHI	0.17	0.31	0.01***
Stage HHI	0.17	0.27	0.01***
Company Region HHI	0.14	0.31	0.01***
Industry Variation	0.56	2.43	0.01***
Stage Variation	0.34	1.76	0.01***
Company Region Variation	0.49	2.79	0.01***

Matrix Metrics: Network-Based Systemic Risk Scoring

Sanjiv Ranjan Das
Leavey School of Business
Santa Clara University
Email: srdas@scu.edu
http://algo.scu.edu/∼sanjivdas/ [1]

September 4, 2014

[1]I am grateful for comments and feedback from Adrian Alter, Ed Altman, Menachem Brenner, Jorge Chan-Lau, Marco Espinosa-Vega, Dale Gray, Levent Guntay, Raman Kapur, Sanjul Saxena, Shann Turnbull, participants at the Consortium for Systemic Risk Analytics MIT; the International Risk Management Conference, Poland; International Monetary Fund; Federal Deposit Insurance Corporation.

Abstract

Matrix Metrics: Network-Based Systemic Risk Scoring

I propose a novel framework for network-based systemic risk measurement and management. A new systemic risk score is defined that depends on the level of risk at each financial institution and the interconnectedness of all banks. This risk metric is decomposable into risk contributions from each entity, forming a basis for taxing each entity appropriately. Risk increments to assess potential risk of each entity are computable. The paper develops many other new risk measures such as system fragility and entity criticality. An assessment using a measure of spillover risk is obtained to determine the scale of externalities that one bank might impose on the system; the metric is robust to this cross risk, and does not induce predatory spillovers. Interestingly, the analysis shows that eliminating too-big-to-fail banks from the system need not lower systemic risk. The new risk metric is contrasted to other metrics in the specific domain of systemic financial risk.

1 Introduction

Morpheus: Unfortunately, no one can be told what the Matrix is. You have to see it for yourself. —The Matrix

The paper proposes a new measure of aggregate systemic risk on networks, and additional system-wide and entity-specific metrics. The measure provides a quantification of system-wide risk based on the level of vulnerability of each node in the system, and the interconnectedness of all nodes in the network (see Alter, Craig, and Raupach (2014) for an approach that also uses the same two quantities). This metric is easy to compute and has many appealing characteristics. Systemic risk (as opposed to systematic risk) has become an important concern after the financial crisis of 2008. Measuring this risk and managing it are two salient goals of the analysis in this paper.

Systemic risk is not always easy to define. But there exist some universally accepted characteristics in the extant literature: a risk that has (a) *large* impact, (b) is *widespread*, i.e., affects a large number of entities or institutions, and (c) has a ripple effect that *endangers the existence* of the financial system. The mortgage/financial crisis of 2008 certainly had all these three characteristics, but the market crash of 1987 impacted only a small set of assets (equities) and did not endanger the financial system. However, definitions of systemic risk abound and economists may not agree on any single one. We describe and compare some popular measures with our new metric.

There is a growing literature on systemic risk measurement in finance, and we mention some representative papers here, though there is a range of papers similar to these. Much of this literature uses equity returns of financial institutions and the correlations of these returns to construct systemic risk measures. An important paper is Billio, Getmansky, Lo, and Pelizzon (2012); they use return correlations and Granger causality regressions on returns to construct network maps and develop network measures of systemic risk. Joint extreme tail risk is also used as a systemic risk measure, such as the well-known *CoVaR* metric of Adrian and Brunnermeier (2010). The *SES* (systemic expected shortfall) measure of Acharya, Pedersen, Philippon, and Richardson (2011) examines tail risk for a financial institution when the aggregate system is under stress. This is similar to the *DIP* (distressed insurance premium) metric of Huang, Zhou, and Zhou (2011).

The systemic risk measure in this paper is different from the ideas in these related papers. First, it does not depend only on equity returns, as it is general and can be used with any measure of interconnectedness. For

example, a network graph generated from interbank transactions may be used as developed in Burdick et al (2011), as might be the network generated from Granger causality analysis in Billio, Getmansky, Lo, and Pelizzon (2012). Second, the measure separates two aspects of overall risk: *compromise* level (i.e., risk score at each node) and *connectivity* (i.e., the network graph) across nodes; it explicitly uses the network matrix in scoring systemic risk. Third, an important property of this aggregate systemic risk measure is that it is decomposable additively into individual contributions to systemic risk, enabling imposing a tax financial institutions, if a regulator so chooses, for individual institutional contributions to aggregate risk.

In addition to these features of the systemic risk score, other useful attributes and applications of this measure are as follows. One, it may be used in combination with network *centrality* scores to manage the risk of the financial system. The *criticality* of a node in the financial system is defined as the product of its risk (compromise) level and its centrality. Two, we propose a measure of *fragility* that is related to concentration risk, i.e., resembles a Herfindahl-Hirschman index. This enables assessment of the speed with which contagion can spread in the system. Three, we compute risk *increments* of the aggregate systemic risk score, i.e., the extent to which each node in the system contributes to aggregate risk if its level of compromise increases by unit amount. This enables identifying which nodes are critical, even though they may not be compromised at the current time. Fourth, we define a *normalized* systemic risk score as well, which quantifies the network effect present in the system. This complements the fragility score. Fifth, we examine *cross risk*, i.e., the externality effect of one node's increase in risk on the risk contribution of other nodes. We explore this risk numerically and find that cross risk is low, i.e., it is not easy for a badly performing node to impose large externalities on the other nodes in terms of our metric, making it a robust one for practical use. Finally, we examine whether breaking large banks into smaller banks helps reduce systemic risk, and find that this remedy does not work. In other words, eliminating too-big-to-fail banks does not eliminate systemic risk.

This short paper proceeds as follows. In Section 2 we present the notation and structure of the new systemic risk score, and we also present related network measures. In Section 3 we extend the measure to a normalized one, and provide more examples. In order to set this metric in context, we provide an extensive Section 4 that summarizes other prominent systemic risk measures, and compares the new metric to these papers. Section 5 provides

brief concluding discussion.

2 Modeling

2.1 Notation

Risk in a connected network arises from compromised nodes and their connections. We propose and define a parsimonious and intuitive metric for the aggregate risk in a network of related entities and explore its properties.

Our network is a graph $G(V, E)$ where $V \in \mathcal{R}^n$ is the vertex (node) set of entities, and $E \in \mathcal{R}^{n \times n}$ is the edge (link) set comprising elements $E(V_i, V_j) \equiv E_{ij} \in \{0, 1\}$. There are n nodes in the network (graph) as indicated above. The graph may be assumed to be directed, i.e., $E_{ij} \neq E_{ji}$, and undirected graphs are special cases. Also, $E_{ii} = 1, \forall i$, and we will see that this is needed for computing the risk score below.

See Figure 1 for an example of the network and matrix. The network is represented by an $(n \times n)$ adjacency matrix with all elements in $\{0, 1\}$. However, one may imagine more complex networks where the connectivity is not binary, but depends on the degree of interaction between nodes. These matrices may be normalized such that the diagonal $E_{ii} = 1, \forall i$, and the off-diagonal elements are scaled to values $E_{ij}/\max(E_{ij}), \forall i, j, i \neq j$. The networks in this paper are not required to be *symmetric* ($E_{ij} = E_{ji}$) or *regular* ($\sum_{i \neq j} E_{ij} = \sum_{j \neq i} E_{ji} = \text{constant}$) as defined in Acemoglu, Ozdaglar, and Tahbaz-Salehi (2013). The set up is simple, yet general.

For each node V_i we define the level of compromise as $C_i \geq 0$. The risk vector for all nodes is $C = [C_1, C_2, ..., C_n]^\top \in \mathcal{R}^n$. Our risk score is agnostic as to how compromise is defined. For example, a good measure of compromise to use would be the Altman (1968) Z-score. Another choice would be the expected loss measure for a financial institution as used in Acharya, Pedersen, Philippon, and Richardson (2011).

2.2 Systemic Risk Score (S)

Definition: The risk score for the entire network is

$$S(C, E) = \sqrt{C^\top E\, C} \tag{1}$$

Scalar S is a function of the compromise level vector C for all nodes and the connections between nodes E. The function $S(C, E)$ is linear homogenous in

C, and this will be shown to be useful in ensuing analytics.

Example: Suppose we have 18 nodes in a network, depicted by the adjacency matrix and directed, unweighted graph in Figure 1. The compromise vector is $C = [0, 0, 1, 2, 2, 2, 2, 2, 1, 0, 2, 2, 2, 2, 1, 0, 1, 1]^\top$, where 0 is no compromise, 1 is a low level of compromise, and 2 indicates a highly compromised node. The risk score using equation (1) is $S = 11.62$.

2.3 Risk Decomposition (D)

Definition: Risk Decomposition is the attribution of the aggregate network risk score S to each node $S_i, i = 1, 2, ..., n$, such that $S = \sum_{i=1}^{n} S_i$.

We exploit the linear homogeneity of the function $S(C, E)$ in C using Euler's equation that decomposes first-order homogenous functions:

$$S = \frac{\partial S}{\partial C_1} C_1 + \frac{\partial S}{\partial C_2} C_2 + \ldots + \frac{\partial S}{\partial C_n} C_n \tag{2}$$

This equation provides a decomposition of the system-wide risk score S into the contribution of each node to the risk. The risk contribution of each node is $D_i = \frac{\partial S}{\partial C_i} C_i$.

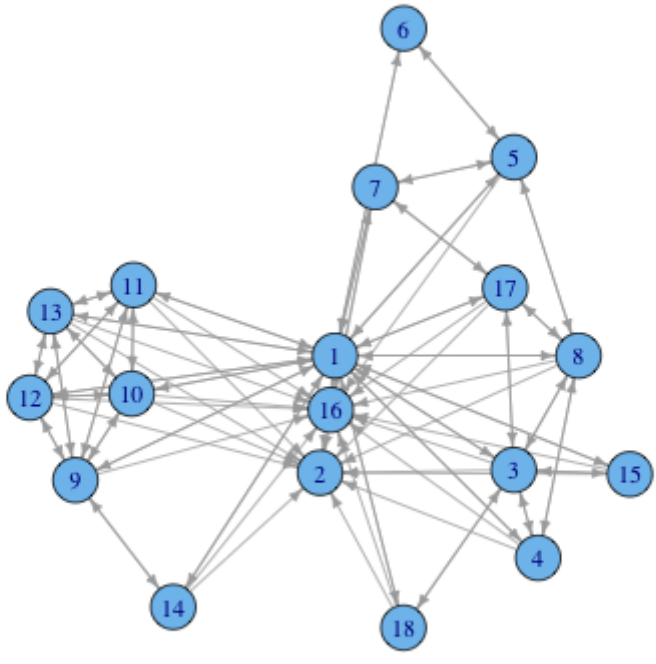
Example: We compute the risk decomposition of the network in Figure 1 and this is shown in Figure 2 where $\sum_{i=1}^{n} D_i = 11.62$. Note that the numbers S_i for each node i depend on both the compromise vector C and the network adjacency matrix E. In this risk network, nodes 5 and 8 contribute the most to system-wide risk.

This risk decomposition is especially useful for pinpointing the network effect when there is a sudden rise in systemic risk score S. By examining the changes in risk contribution for each node, the causal node is quickly identified.

2.4 Risk Increment (I)

Definition: Risk Increment is the change in the aggregate network risk score S when the compromise score c_i of an asset changes, i.e., $I_i = \frac{\partial S}{\partial C_i}$.

Example: We compute the risk increments of the network in Figure 1 and this is shown in Figure 3. Note that the numbers I_i for each node i depend on both the compromise vector C and the network adjacency matrix E.



	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]	[,13]	[,14]	[,15]	[,16]	[,17]	[,18]
[1,]	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
[2,]	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
[3,]	1	1	1	1	0	0	0	1	0	0	0	0	0	0	1	1	1	1
[4,]	1	1	1	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0
[5,]	1	0	0	0	1	1	1	1	0	0	0	0	0	0	0	1	0	0
[6,]	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
[7,]	1	1	0	0	1	0	1	0	0	0	0	0	0	0	0	1	1	0
[8,]	1	1	1	1	1	0	0	1	0	0	0	0	0	0	0	1	1	0
[9,]	1	0	0	0	0	0	0	0	1	1	1	1	1	1	0	1	0	0
[10,]	1	1	0	0	0	0	0	0	1	1	1	1	1	0	0	1	0	0
[11,]	1	1	0	0	0	0	0	0	1	1	1	1	1	0	0	1	0	0
[12,]	1	1	0	0	0	0	0	0	1	1	1	1	1	0	0	1	0	0
[13,]	1	1	0	0	0	0	0	0	1	1	1	1	1	0	0	1	0	0
[14,]	1	1	0	0	0	0	0	0	1	0	0	0	0	1	0	1	0	0
[15,]	1	1	1	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0
[16,]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0
[17,]	1	1	1	0	0	0	1	1	0	0	0	0	0	0	0	1	1	0
[18,]	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1

Figure 1: Directed network of 18 nodes. One-way arrows means that risk flows in the direction of the arrow. Two-way arrows means risk flows in both directions. The network is summarized in the adjacency matrix. Note that the diagonal values are all 1. The "diameter" of this network, i.e., the maximal shortest distance between any two nodes is 2.

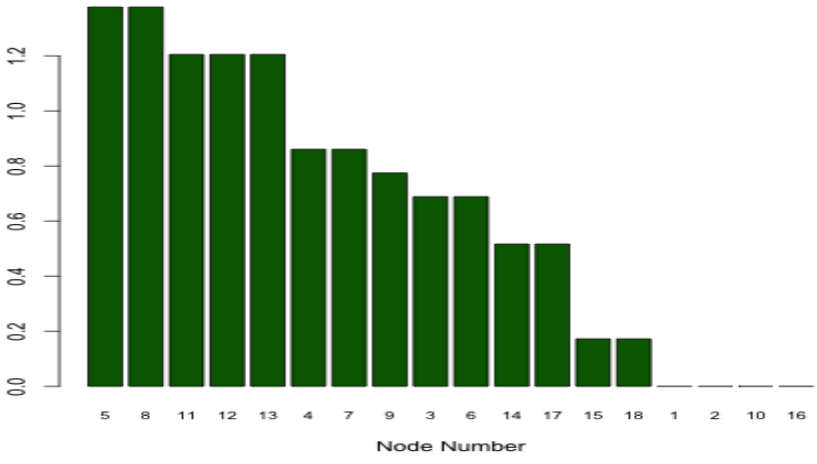


Figure 2: *Risk Decomposition*: The risk contribution S_i for each node in the network shown in Figure 1, rank ordered for display. The aggregate risk is $\sum_{i=1}^{20} S_i = 11.62$.

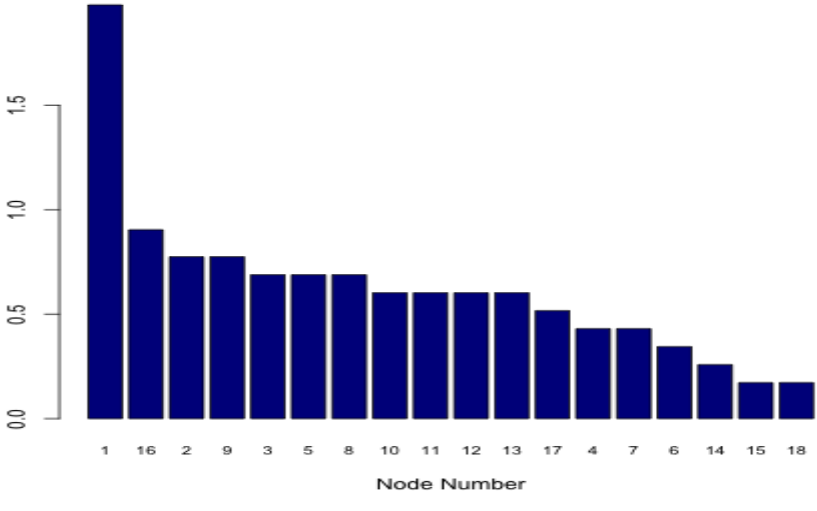


Figure 3: *Risk Increment*: I_i for each node in the network shown in Figure 1, rank ordered for display.

Both, risk contribution and risk increment are useful in identifying the source of system vulnerabilities, and in remediation. In assessing whether a node should be allowed to fail, we may disconnect it from the network and assess how these metrics are impacted. Note that node 1 has a very low current risk contribution (as shown in Figure 2), but has the potential to be very risky as it has the highest risk increment (see Figure 3).

2.5 Fragility (R)

Definition: We define the "fragility" of the network $R = E(d^2)/E(d)$, where d is the degree of a node, i.e., the number of connections it has to other nodes.

This definition is intuitive, and is a measure that is similar to a normalized Herfindahl-Hirschman index. If the network's connections are concentrated in a few nodes, we get a hub-and-spoke network (also known as a scale-free network) on which spread of a shock is rapid, because once a node with many connections is infected, disease on a network spreads rapidly. Consider for example a network with four nodes each with degree 2, a network that is not fragile, i.e., fragility score is low, $R = 2$, but the same network of four nodes with degrees $\{4, 2, 1, 1\}$ has the same mean degree, but is much more fragile as $R = 11$. Concentration of degree induces fragility. This metric is a useful complement to the systemic risk score S. The fragility of the example network is computed to be 7.94.

2.6 Centrality (x) and Criticality (y)

Definition: Eigenvalue "centrality" is the normalized principal eigenvector $x \in \mathcal{R}^n$ such that for scalar λ, satisfies the eigensystem

$$\lambda\, x = E\, x \tag{3}$$

Centrality was first defined in Bonacich (1987), and popularized more recently as the PageRank algorithm by Google [Brin and Page (1998)].

Example: We compute centrality for this network and plot it in Figure 4. Note that neither centrality or fragility depend on the vector C. Centrality is normalized where the highest centrality node is set to value 1, and the other node values are relative centrality to this node.

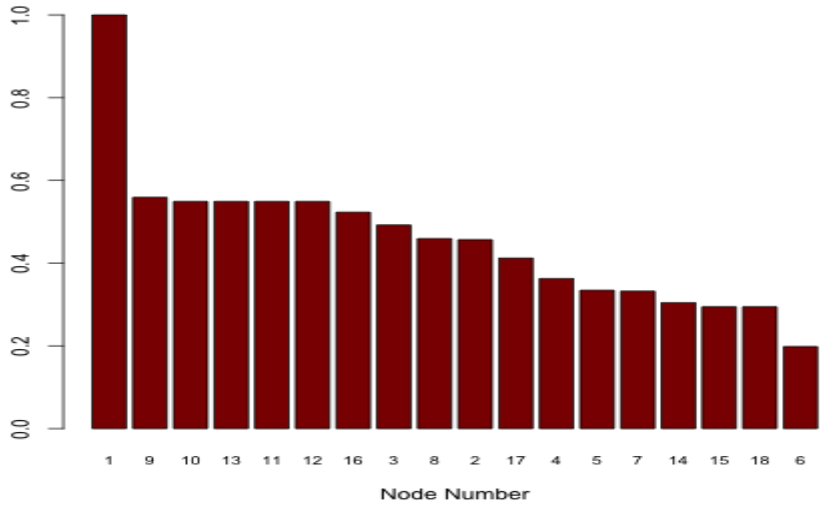


Figure 4: *Centrality*: Normalized centrality for each node in the network shown in Figure 1, rank ordered for display.

Definition: "Criticality" is compromise-weighted centrality. This new measure is defined as $y = C \times x$ where $y, C, x \in \mathcal{R}^n$. Note that this is an element-wise multiplication of vectors C and x.

Critical nodes need immediate attention, either because they are heavily compromised or they are of high centrality, or both. It offers a way for regulators to prioritize their attention to critical financial institutions, and pre-empt systemic risk from blowing up. We compute criticality for this network and plot it in Figure 5.

3 Extended Metrics

The previous section introduced several new network-based systemic risk measures such as the aggregate systemic risk score, risk decomposition, risk increment, fragility, and criticality. In this section, we modify and extend these metrics further.

The units of systemic risk score S are determined by the units of compromise vector C. If C is a rating, then systemic risk S is measured in rating

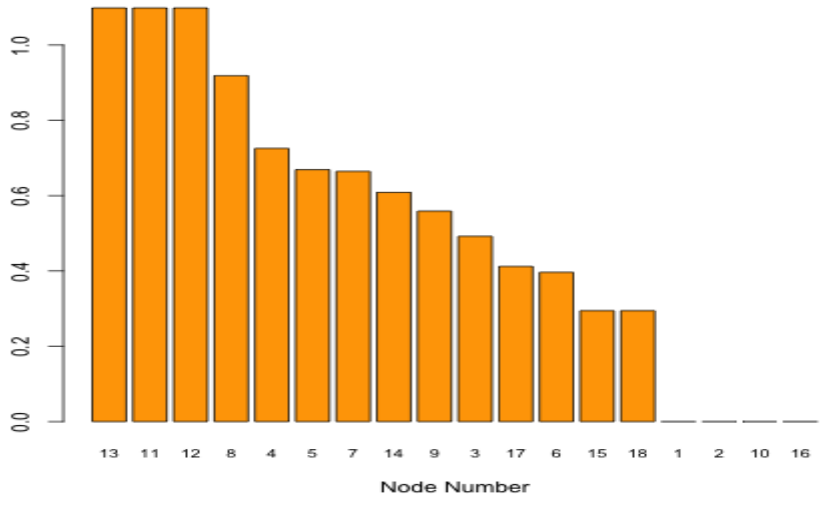


Figure 5: *Criticality*: Criticality for each node in the network shown in Figure 1, rank ordered for display.

units. If C is a Z-score (for instance), then S is a system-wide Z-score. And if C is expected loss, then S is in system-wide expected loss units.

3.1 Normalized Risk Score ($\bar{S}$)

In order to compare the systemic risk score across systems, we extend the score S to normalized score $\bar{S}$:

$$\bar{S} = \frac{\sqrt{C^\top E\, C}}{\|C\|} = \frac{S}{\|C\|} \tag{4}$$

where $\|C\| = \sqrt{C^\top C}$ is the norm of vector C. When there are no network effects, $E = I$, the identity matrix, and $\bar{S} = 1$, i.e., the normalized baseline risk level with no network (system-wide) effects is unity. We can use this normalized score to order systems by systemic risk. For the system in our example, the normalized score is $\bar{S} = 1.81$.

3.2 Varying Risk or Connectivity

The normalized score $\bar{S}$ has intuitive properties. For example, the addition of a link in the network will increase $\bar{S}$ ceteris paribus. And a reallocation of risk among nodes in vector C will also change $\bar{S}$. Limiting entries in matrix E is akin to controls on counterparty risk in an interbank system, and limiting each entry in vector C constrains own risk. A network regulator may choose limits in different ways to manage systemic risk. Simulating changes to C and E allows for generating interesting test case scenarios of systemic risk.

Example: (*Increasing risk at a node*) If we keep the example network unchanged, but re-allocate the compromise vector by reducing the risk of node 3 by 1, and increasing that of node 16 by 1, we find that the risk score S goes from 11.62 to 11.87, and the normalized risk score $\bar{S}$ goes from 1.81 to 1.85.

Example: (*Increasing linkages in the system)* Suppose, in the example network, we add one additional bi-directed link between nodes 6 and 12 (see Figure 1). The risk score S increases from 11.62 to 11.96, and the normalized risk score $\bar{S}$ increases from 1.81 to 1.87.

Thus, we may examine how adding a link to the network or removing a link may help in reducing system-wide risk. Or we may examine how additional risk at any node leads to more systemic risk. A system regulator can run these analyses to determine the best way to keep system-wide risk in check.

3.3 Cross/Spillover Risk (ΔD_{ij})

An increase in the risk level at any node does not only impact its own risk contribution, but that of other nodes as well. A single financial institution mismanaging its own risk might impose severe externalities in terms of potential risk on other banks in the system through network effects. In a situation where banks are taxed for their systemic risk contributions, for example, required to keep additional capital based on their individual risk contributions (D_i), externalities may instigate retaliatory actions that result in escalation in systemic score S. Hence, it is important to compute how severe cross risk

might be. Espinosa-Vega and Sole (2010) point out that spillover risk is an important motivation for their model of capital surcharges for systemic risk in their model of financial surveillance.

We analyzed our sample network by computing the effect on risk contribution of each node if any other node has a unit increase in compromise level. We denote the cross risk of node i when node j has a unit increase in compromise level C_j as $\Delta D_{ij} = \frac{\partial D_i}{\partial C_j}$. The results are shown in Figure 6. It is apparent that cross risk is insignificant compared to own risk contribution. This suggests that regulators need not be overly concerned with moral hazard on networks, where one node can impose severe externalities on other nodes.

3.4 Risk Scaling

We asses three questions here, in order to derive a deeper understanding of the properties of systemic risk score S. These questions pertain to how the score changes when we scale the level of compromise, the level of interconnectedness, and the breaking down of nodes into less connected ones.

First, ceteris paribus, how does an across the board change in compromise vector C impact S? We note that since S is linear homogenous in C, this effect is purely linear.

Second, how does an increase in connectivity impact $\bar{S}$? Is this a linear or non-linear effect? We ran a simulation of a fifty node network and examined $\bar{S}$ as the number of connections per node was increased. Simulation results are shown in Figure 7. The plot shows how the risk score increases as the probability of two nodes being bilaterally connected increases from 5% to 50%. For each level of bilateral probability a random network is generated for 50 nodes. A compromise vector is also generated with equally likely values $\{0, 1, 2\}$. This is repeated 100 times and the mean risk score across 100 simulations is plotted on the y-axis against the bilateral probability on the x-axis. These results based on random graph generation show that the risk score increases with connectivity, but in less than linear fashion (the plot is concave). This corresponds to results in Vivier-Lirimont (2006); Blume, Easley, Kleinberg, Kleinberg, and Tardos (2011); Gai, Haldane, and Kapadia (2011) who show that dense interconnections destabilize networks.

Third, we examine whether partitioning nodes into more numerous smaller entities reduces systemic risk (a question also addressed in very different models by Cabrales, Gottardi, and Vega-Redondo (2014); Vivier-Lirimont

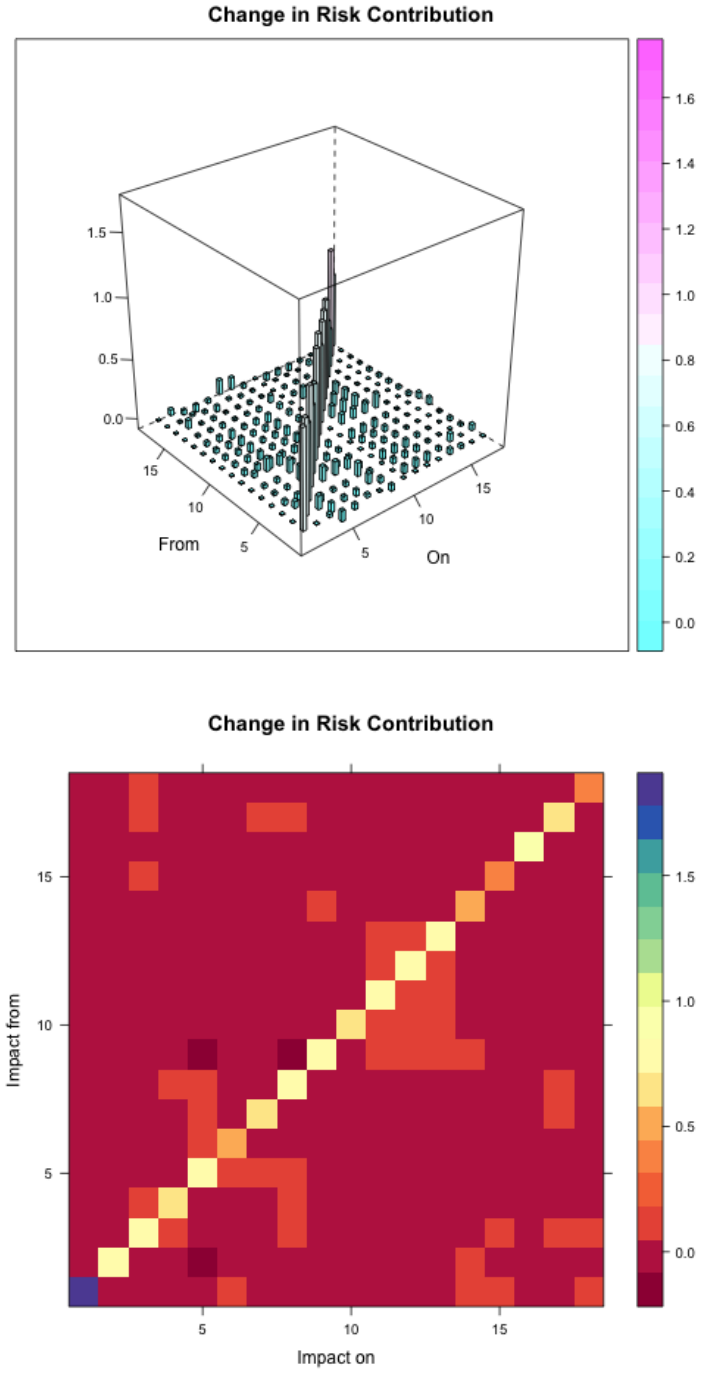


Figure 6: Change in risk contribution when any node experiences a unit increase in compromise level. The impact from each node on every other node is shown. The upper plot is in bar form, and the lower plot is a heat map.

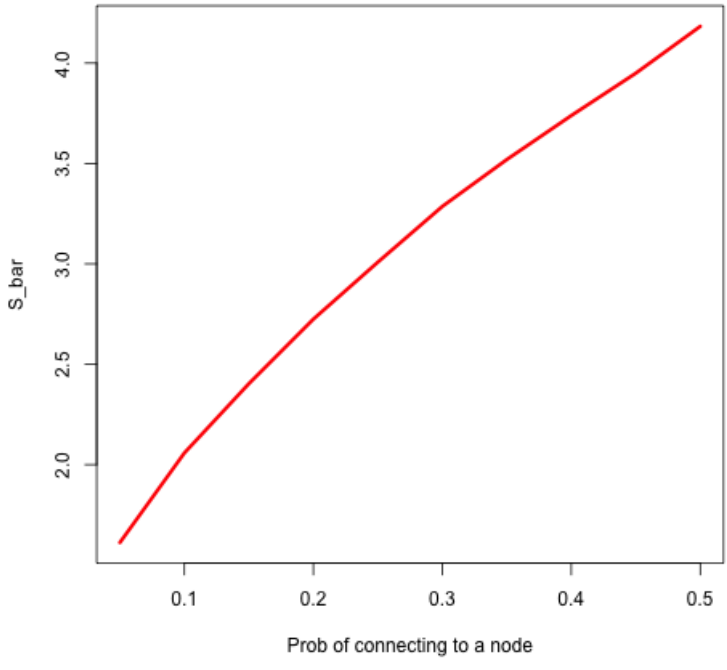


Figure 7: The increase in normalized risk score $\bar{S}$ as the number of connections per node increases. The plot shows how the risk score increases as the probability of two nodes being bilaterally connected increases from 5% to 50%. For each level of bilateral probability a random network is generated for 50 nodes. A compromise vector is also generated with equally likely values $\{0, 1, 2\}$. This is repeated 100 times and the mean risk score across 100 simulations is plotted on the y-axis against the bilateral probability on the x-axis.

(2006)). Whereas the first two questions did not consider increasing or decreasing the number of nodes, in this case we explicitly increase the numbers of nodes while adjusting the average number of connections per node down, so as to keep the overall connectivity unchanged, while changing the structure of the network. Figure 8 shows that the risk score $\bar{S}$ remains materially unaffected for all practical purposes, hence, splitting large banks into smaller banks does not reduce systemic risk. This risk does not require the presence of too-big-to-fail banks.

4 Comparison with Other Measures of Systemic Risk

As a practical matter, several measures of systemic risk have been proposed, and each one implicitly defines systemic risk as that risk being quantitatively determined by their measure. This is definition by quantification, measurement as one sees it. In our setting of risk networks the system-wide risk scores $\{S, \bar{S}\}$ capture systemic risk as a function of the compromise vector C and the network of connected risk entities E. Other research conducts this differently. Some measures of systemic risk are network-based but most of the measures are based on stock return correlations.

1. Billio, Getmansky, Lo, and Pelizzon (2012) define two measures of systemic risk across banks, hedge funds, broker/dealers, and insurance companies. The idea is to measure correlations among institutions directly and unconditionally using principal components analysis and Granger causality regressions, and thereby assess the degree of connectedness in the financial system.

   In their framework, total risk of the system is the variance of the sum of all financial institution returns, denoted σ_S^2. PCA comprises an eigenvalue decomposition of the covariance matrix of returns of the financial institutions, and systemic risk is higher when the number of principal components n that explain more than a threshold H of the variation in the system is small. Using notation in their paper,

   $$h_n = \frac{\omega_n}{\Omega} > H, \tag{5}$$

   where h_n is the fraction of σ_S^2 that is explained by the first n components, i.e., $\Omega = \sum_{i=1}^{N} \lambda_i$ and $\omega_n = \sum_{i=1}^{n} \lambda_i$, where λ_i is the i-th eigenvalue. We note that σ_S is linear homogenous, so can be decomposed to

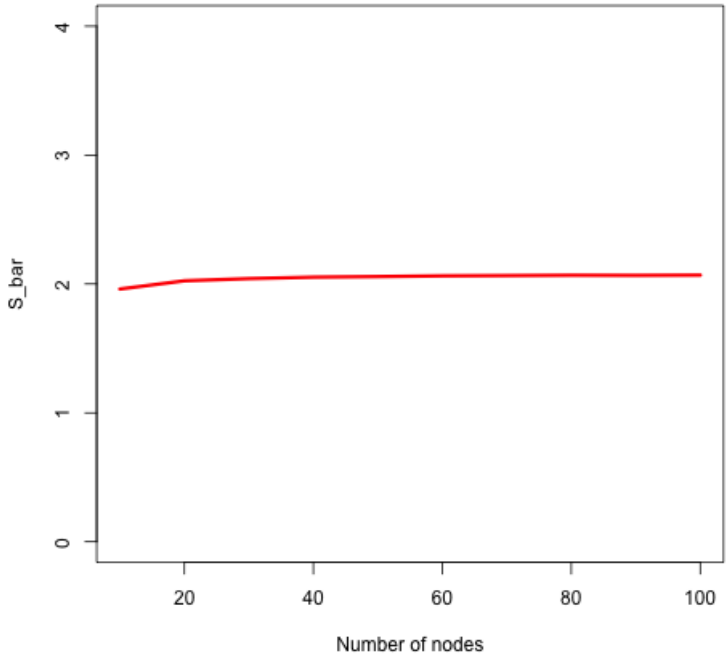


Figure 8: The change in normalized risk score $\bar{S}$ as the number of nodes increases, while keeping the average number of connections between nodes constant. This mimics the case where banks are divided into smaller banks, each of which then contains part of the transacting volume of the previous bank. The plot shows how the risk score increases as the number of nodes increases from 10 to 100, while expected number of total edges in the network remains the same. A compromise vector is also generated with equally likely values $\{0, 1, 2\}$. This is repeated 5000 times for each fixed number of nodes and the mean risk score across 5000 simulations is plotted on the y-axis against the number of nodes on the x-axis.

obtain the risk contribution of each financial institution, in the same manner as is done for our network risk measure S.

In addition to this covariance matrix based measure of systemic risk, Billio, Getmansky, Lo, and Pelizzon (2012) also create a network using Granger causality. This directed network is represented by an adjacency matrix of values $(0, 1)$ where node i connects to node j if the returns of bank i Granger cause (in a linear or nonlinear way) those of bank j, i.e., edge $E_{i,j} = 1$. This adjacency matrix is then used to compute connectedness measures of risk such as number of connections, fraction of connections, centrality, and closeness. These measures correspond to some of those presented in the exposition above, and the first two report an aggregate measure of system-wide risk, different from the S measure developed in this paper. Again, since system-wide risk is defined as a count of the number of connections, it is easy to determine what fraction is ascribable to any single financial firm. They applied the metrics to U.S. financial institution stock return data, and in a follow-up paper, to CDS spread data from U.S., Europe, and Japan (see (Billio, Getmansky, Gray, Lo, Merton, and Pelizzon, 2014)), where the global system is also found to be highly interconnected.

Overall, we note a strong complementarity between the analyses in Billio, Getmansky, Lo, and Pelizzon (2012) and our paper, and using the network matrix in their paper, we may implement our systemic risk score S as well. Hence, this paper extends and uses the results in this earlier work.

2. The $CoVaR$ measure of Adrian and Brunnermeier (2010) estimates a bank or the financial sector's Value at Risk (VaR) given that a particular bank has breached its VaR. They use quantile regressions on asset returns (R) using data on market equity and book value of debt. Pairwise $CoVaR(j|i)$ for bank j given bank i is at VaR is defined implicitly as the quantile α satisfying

$$Pr[R_j \leq -CoVaR_\alpha(j|i)|R_i = -VaR_\alpha(i)] = \alpha \tag{6}$$

where $VaR(i)$ is also defined implicitly as $Pr[R_i \leq -VaR_\alpha(i)] = \alpha$. The actual measure of systemic risk is then

$$\Delta CoVaR_\alpha(j|i) = CoVaR_\alpha(j|i) - VaR_\alpha(j) \tag{7}$$

The intuition here is one of under-capitalization when a systemic event occurs, i.e., extra capital needed because capital needed for solvency at the time of a systemic event ($CoVaR_\alpha(j|i)$) is greater than capital needed in normal times ($VaR_\alpha(j)$). Replacing j with the system's value $\Delta CoVaR_\alpha(S|i)$ gives an aggregate measure of systemic risk. However, this is still not an aggregate measure of risk (such as S in this paper), rather one that assesses the systemic risk increment or contribution of the i-th financial institution.

3. The SES (systemic expected shortfall) measure of Acharya, Pedersen, Philippon, and Richardson (2011) captures the amount by which an otherwise appropriately capitalized bank is undercapitalized in the event of a systemic crisis. It is related to MES (marginal expected shortfall), which is the average return of a financial institution for the 5% worst days in the market. Mostly, SES is analogous to $CoVaR$ where Value-at-Risk is replaced with expected shortfall (ES), though the implementation details and variables used differ in the paper of Acharya, et al. We may write think of SES as the equity shortfall a firm experiences when aggregate banking equity $e(S)$ is below a threshold H, i.e.,

$$SES(j) = E[H(j) - e(j)|e(S) \leq H] \tag{8}$$

where $H(j)$ is the desired threshold level of equity for bank j, with equity level $e(j)$. SES has useful properties in that it is in dollar terms and scales with institution size, so that is it easily aggregated. The DIP (distressed insurance premium) measure of Huang, Zhou, and Zhou (2011) is similar to the SES of Acharya, Pedersen, Philippon, and Richardson (2011) in that it also captures the expected losses of a financial institution conditional on losses being greater than a threshold level.

There is an important difference between the between the Granger causality based network of Billio, Getmansky, Lo, and Pelizzon (2012) and $CoVaR$, versus the SES measure. The two former measures assess the impact a single bank has on the system, whereas the latter measure assesses the impact of system-wide risk on each bank. The new measures of system-wide risk $(S, \bar{S})$ proposed in this paper are akin to the first approach, and I believe that this is the more relevant view of systemic risk, and offers an aggregate risk score

as well. However, both approaches are relevant in computing extra systemic risk capital requirements.

There are some important differences between these measures of systemic risk and the network score in this paper.

1. These measures focus on the effect of failure of a given institution on others. Hence, they are pairwise and conditional. In contrast, network risk scores are system-wide and unconditional.

2. The measures are based on correlations, and correlations tend to be high in crisis periods but are not early-warning indicators of systemic risk. Relying on stock return correlations as an early warning indicator of network risk is likely to be futile, as correlation matrices reflect systemic risk after the risk has arisen, rather than before. network-based measures may be better at identifying if there is a systemic vulnerability prior to a system shock.

3. Correlation based measures tend to be removed from the underlying mechanics of the system, and are in the nature of implicit statistical metrics. network-based measures directly model the underlying mechanics of the system because the adjacency matrix E is developed based on physical transaction activity. Further, the compromise vector is a function of firm quality that may be measured in multidimensional ways. This separation of network effect (connectivity) and individual bank risk (compromise), and their combination into a single aggregate risk score, offers a simple, practical, and general approach to measuring systemic risk.

This paper is not only related to the growing literature on measures of systemic risk, but also to the network literature in economics in papers like Acemoglu, Ozdaglar, and Tahbaz-Salehi (2013); Allen and Gale (2000); Allen, Babus, and Carletti (2012), and the literature on risk in clearing systems, see Eisenberg and Noe (2001); Duffie and Zhu (2011), Borovkova and El Mouttalibi (2013). Systemic risk measures based on dynamic conditional correlations are also proposed, see Brownlees and Engle (2010); Engle, Jondeau, and Rockinger (2012).

Therefore, the novel framework in this paper may be used as a complement to existing approaches. Whether or not the network is derived from

physical deal flow or from returns data, the risk score S may be computed, decomposed by node, and risk increments derived therefrom, along with many other metrics, to provide a useful dashboard for managing systemic risk.

5 Concluding Comments

This framework for network-based systemic risk modeling develops system-wide risk scores such as a new aggregate systemic risk score (S), a normalized score ($\bar{S}$), a fragility score (R), and also entity-specific risk scores: a risk decomposition (D_i), risk increments (I_i), criticality (y_i), and a score for spillover risk (ΔD_{ij}). All these metrics use simple data inputs: an institution specific compromise vector C, and the network graph of financial institution linkages E. The risk metrics are general, i.e., independent of the particular definitions of C, E, and also complement and extend systemic risk measures in the extant literature.

Modeling extensions are also envisaged. In the current version of the model the compromise vector C is independent of the connectivity matrix E. Making C a function of E (and vice versa) leads to interesting additional implications, and of course, fresh econometric questions. For example, C may be an increasing function of E, but then E may be a decreasing function of C, making it unclear as to whether an increase in risk or transaction volume always leads to leads to a higher level of potential systemic risk. Issues such as the structure of the network and the interaction of its components are addressed in the models of Allen, Babus, and Carletti (2012); Glasserman and Young (2013); Elliott, Golub, and Jackson (2014). The welfare implications of over linking are discussed in the contagion model of Blume, Easley, Kleinberg, Kleinberg, and Tardos (2011).

How to construct composite connectivity matrices across markets is also an interesting issue. One may get a network matrix from transactions in the CDS market (for example) and another from the bond markets, but the question of putting these two matrices (call them E_1 and E_2) together into one composite E matrix requires a weighting scheme or other collapsing technical condition. One solution to this would be to make E the matrix of bilateral CVA (credit valuation adjustment) numbers, because this directly measures the exposure of each financial institution to another across all products and asset classes. Using counterparty exposures as a device is also considered in the "10-by-10-by-10" systemic risk measurement approach recommended in

Duffie (2011).

From a regulatory point of view, there are many applications for this framework. First, imposition of additional capital required may be based on a composite score computed from risk decomposition numbers, taking into account additional informative metrics such as criticality, risk increments, and spillover risk. (See a proposal for this in Espinosa-Vega and Sole (2010).) Second, this composite score may be used to allocate supervision money across various financial institutions. Third, the systemic score can be tracked over time, and empirical work will be needed to backtest whether the systemic score S is a useful early warning predictor of systemic risk events. Using a different approach, Kritzman, Li, Page, and Rigobon (2010); Reyngold, Shnyra, and Stein (2013) find predictability of systemic risk. Fourth, an analysis of network robustness in addition to measuring systemic risk is a complementary analysis, for example Allen and Gale (2000); Callaway, Newman, Strogatz, and Watts (2000).

The poem "No Man Is An Island" by John Donne is metaphorical summary of the ideas and issues discussed in this paper, so it is only apt to reproduce it here:

No man is an island,
Entire of itself,
Every man is a piece of the continent,
A part of the main.
If a clod be washed away by the sea,
Europe is the less.
As well as if a promontory were.
As well as if a manor of thy friend's
Or of thine own were:
Any man's death diminishes me,
Because I am involved in mankind,
And therefore never send to know for whom the bell tolls;
It tolls for thee.

References

Acemoglu, D., Ozdaglar, A., Tahbaz-Salehi, A. (2013). "Systemic Risk and Stability in Financial Networks," Working paper, MIT.

Acharya, V., Pedersen,L.H., Philippon,T., Richardson,M., (2011). "Measuring systemic risk, " Working paper, New York University.

Adrian, T., Brunnermeier, M., (2010). "CoVaR," Working paper, Princeton University.

Allen, F., Gale, D., (2000). "Financial Contagion," *Journal of Political Economy* 108(1), 1–32.

Allen, F., Babus, A., Carletti, E, (2012). "Asset commonality, debt maturity and systemic risk," *Journal of Financial Economics* 104, 519–534.

Alter, A., Craig, B., Raupach, P., (2014). "Centrality-Based Capital Allocations and Bailout Funds," *IMF Working Paper*.

Altman, Edward I. (1968). "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy," *Journal of Finance* 23(4), 189–209.

Billio, M., Getmansky, M., Lo. A., Pelizzon, L. (2012). "Econometric Measures of Connectedness and Systemic Risk in the Finance and Insurance Sectors," *Journal of Financial Economics* 104(3), 536–559.

Billio, M., Getmansky, M., Gray, D., Lo. A., Merton, R., Pelizzon, L. (2012). "Sovereign, Bank and Insurance Credit Spreads: Connectedness and System Networks," Working paper, IMF.

Larry, B., Easley, D., Kleinberg, J., Kleinberg, R., Tardos, E. (2011), "Network formation in the presence of contagious risk," *Proceedings of the 12th ACM Conference on Electronic Commerce*.

Bonacich, P. (1987). "Power and Centrality: A Family of Measures," *American Journal of Sociology* 92(5), 1170–1182.

Borovkova, S., El Mouttalibi, H.L., (2013). "Systemic Risk and Centralized Clearing of OTC derivatives: A Network Approach," Working Paper, VU Amsterdam.

Brin, S., Page, L. (1998). "The anatomy of a large-scale hypertextual web search engine," *Computer Networks and ISDN Systems* 30, 107–117.

Brownless, C.T., Engle, R.F., (2010). "Volatility, Correlation, and Tails for Systemic Risk Measurement," Working Paper NYU, SSRN 1611229.

Burdick, D., Das, S., Hernandez, M. A., Ho, H., Koutrika, G., Krishnamurthy, R., Popa, L., Stanoi, I., Vaithyanathan, S., (2011). "Extracting, Linking and Integrating Data from Public Sources: A Financial Case Study," *IEEE Data Engineering Bulletin* 34(3), 60–67.

Cabrales, A., Gottardi, P., Vega-Redondo, F., (2014). "Risk-sharing and Contagion in Networks," Working paper, University College London.

Callaway, D.S., Newman, M.J., Strogatz, S.H., Watts, D.J., (2000). "Network Robustness and Fragility: Percolation in Random Graphs," *Physical Review Letters* 85(25), 5468–5471.

Duffie, D., Zhu, X., (2011). "Does a Central Clearing Counterparty Reduce Counterparty Risk?" *Review of Asset Pricing Studies* 1, 74–95.

Duffie, D., (2011). "Systemic Risk Exposures: A 10-by-10-by-10 Approach," forthcoming in *Systemic Risk and Macro Modeling*, Markus K. Brunnermeier and Arvind Krishnamurthy, editors, University of Chicago Press.

Eisenberg, L., Noe, T., (2001). "Systemic Risk in Financial Systems," *Management Science* 47(2), 236–249.

Elliott, M., Golub, B., Jackson, M., (2014). "Financial Networks and Contagion," forthcoming *American Economic Review*.

Engle, R., Jondeau, E., Rockinger, M., (2012). "Dynamic Conditional Beta and Systemic Risk in Europe," Working Paper NYU.

Espinosa-Vega, M., Sole, J., (2010). "Cross-Border Financial Surveillance: A Network Perspective," *IMF Working Paper*, WP/10/105.

Gai, P., Kapadia, S. (2010). "Contagion in Financial Networks," Working paper, Bank of England.

Gai, P., Haldane, A., Kapadia, S. (2011), "Complexity, concentration and contagion," *Journal of Monetary Economics* 58, 453–470.

Glasserman, P., Young, H.P., (2013). "How Likely is Contagion in Financial Networks?" Working Paper #0009, Columbia University.

Huang, X., Zhou, Hao., Zhu, H., (2011). "Systemic Risk Contributions," Working paper, Divisions of Research & Statistics and Monetary Affairs, Federal Reserve Board, Washington, D.C.

Kritzman, M., Li, Y., Page, S., Rigobon, R., (2010). "Principal Components as a Measure of Systemic Risk," Working paper, MIT, SSRN 1582687.

Reyngold, A., Shnyra, K., Stein, R., (2013). "Aggregate and Firm-level Measures of Systemic Risk from a Structural Model of Default," MIT LFE Working Paper No. LFE-0501-13.

Vivier-Lirimont, S. (2006), "Contagion in interbank debt networks," Working paper, Reims Management School and CES, Paris I Pantheon Sorbonne University.

Unleashing the Power of Public Data for Financial Risk Measurement, Regulation, and Governance

Mauricio A. Hernández, Howard Ho, Georgia Koutrika, Rajasekar Krishnamurthy, Lucian Popa, Ioana R. Stanoi, Shivakumar Vaithyanathan, Sanjiv Das*

IBM Research – Almaden
{mauricio,ho,lucian,shiv}_at_almaden.ibm.com
{gkoutri,rajase,irs}_at_us.ibm.com

*Finance Department
Leavey School of Business
Santa Clara University
srdas_at_scu.edu

ABSTRACT

We present Midas, a system that uses complex data processing to extract and aggregate facts from a large collection of structured and unstructured documents into a set of unified, clean entities and relationships. Midas focuses on data for financial companies and is based on periodic filings with the U.S. Securities and Exchange Commission (SEC) and Federal Deposit Insurance Corporation (FDIC). We show that, by using data aggregated by Midas, we can provide valuable insights about financial institutions either at the whole system level or at the individual company level. To illustrate, we show how co-lending relationships that are extracted and aggregated from SEC text filings can be used to construct a network of the major financial institutions. Centrality computations on this network enable us to identify critical hub banks for monitoring systemic risk. Financial analysts or regulators can further drill down into individual companies and visualize aggregated financial data as well as relationships with other companies or people (e.g., officers or directors). The key technology components that we implemented in Midas and that enable the above applications are: information extraction, entity resolution, mapping and fusion, all on top of a scalable infrastructure based on Hadoop.

1. INTRODUCTION

During the last few years, we have observed an explosion in the number and variety of public data sources that are available on the web: research papers and citations data (e.g., Cora, Citeseer, DBLP), online movie databases (e.g., IMDB), etc. While many of these sources have been used and studied in recent years by computer science papers, there are, however, other types of public data covering additional domains. Two such significant domains are the business/financial domain and the government/regulatory domain. Examples of business/financial data include company filings with regulatory bodies such as SEC and FDIC, security market (e.g., stock, fund, option) trading data, and news articles, analyst reports, etc. Examples of government data include US federal government spending data, earmarks data, congress data, census data, etc. Yet another domain of significant importance is healthcare.

Public data sources tend to be distributed over multiple web sites, and their contents vary from unstructured (or text) to semi-structured (html, XML, csv) and structured (e.g., tables). In this paper, we will focus on business data sources in the financial domain, with particular emphasis on the filings that companies are required to submit periodically to SEC and FDIC. This allows us to access high-quality (i.e., fresh and post-audit) content that is often cleaner and more complete than community-contributed data sources, e.g., Wikipedia. Nevertheless, even though highly regulated, the SEC and FDIC data still poses challenges in that a large number of filings are in text. Thus, to extract and integrate key concepts from SEC filings, information extraction technology becomes a crucial part in the overall data flow.

In this paper, we present our experience with building and applying Midas, a system that unleashes the value of information archived by SEC and FDIC, by extracting, conceptualizing, integrating, and aggregating data from semi-structured or text filings. We show that, by focusing on high-quality financial data sources and by combining three complementary technology components – information extraction, information integration, and scalable infrastructure – we can provide valuable insights about financial institutions either at the whole system level (i.e., systemic analysis) or at the individual company level. A major step towards providing such insights is the aggregation of fine-grained data or facts from hundreds of thousands of documents into a set of clean, unified entities (e.g., companies, key people, loans, securities) and their relationships. In other words, we start from a document-centric archive, as provided by SEC and FDIC, and build a concept-centric repository (a "Web of Concepts" [10]) for the financial domain that enables sophisticated structured analysis.

We exhibit two types of financial applications that can be built on top of our consolidated data. First, we show how we can construct a network of the major financial institutions where the relationships are based on their aggregated lending and co-lending activities. By employing centrality computation, we show that a few major banks (J P Morgan Chase & Co, Citigroup Inc, Bank of America) are critical hubs in the network, as they have high connectivity to all the important components in the network. Hence, their systemic risk is high. While the results are intuitively as expected, they show that our data-driven analysis can lead to accurate results even by employing a few key relationships (in this case, just co-lending). The second type of applica-

WWW2010, April 26-30, 2010, Raleigh, North Carolina.

tion is the drill-down inside the individual aggregated entities. For example, if Citigroup is identified as a critical hub in the global network, regulators may wish to drill down into the various aspects related to Citigroup. To this extent, we provide multiple aggregated views that include:

- the list of key executives or insiders (either officers or directors), with their full employment history (including the movement across companies);
- the transactions (e.g., stock buys or sells) that insiders make, and the general trends of such insider transactions. As an example, having more buys than sells in a year may indicate either a strong company or simply that the market is at a low point;
- the relationships (of a given company) to other companies; this includes identifying subsidiaries of a company, institutional holdings in other companies, potential competitors based on movement of executives, as well as companies that are related via lending/borrowing activities.

These views foster tracking senior executives, and company interrelationships, etc., that are key components of monitoring corporate governance in financial institutions.

Midas employs a number of scalable technology components to achieve the desired level of integration. All components can process large number of documents and run as map/reduce jobs on top of Hadoop. One component is in charge of information extraction from unstructured sources and is based on SystemT [14]. This component includes high-level rules (expressed in AQL, the SystemT language) to extract structured data from unstructured text. The rest of the components are in charge of the structured information integration. Essentially, these components map and merge the extracted data into a pre-defined schema (e.g., Person). An *entity resolution* component helps identify references to the same real-world entity across the multiple input documents. All these components are implemented in Jaql [3], a high-level general language that compiles data transformations as Hadoop jobs.

This paper is organized as follows. Section 2 details some of the complex analysis that Midas enables. Section 3 explains the components in the Midas integration flow and Section 4 describes the public data sources that we used. Section 5 then explains how we programmed Midas to extract and integrate data from these public data sources. We conclude in Section 6 with an outlook of other applications that can benefit from Midas technology.

2. MIDAS: THE APPLICATIONS

In this section, we discuss the types of financial applications that the data aggregated by Midas enables. We group these applications into two types (one systemic, and one at the individual company level).

2.1 Systemic Risk Analysis

"Systemic" effects have emerged as a leading concern of economic regulators in the past few years since the financial crisis began in 2007/2008. Recessionary conditions result, of course, in the failure of individual financial institutions, but systemic risk is primarily concerned with the domino effect of one financial institution's failure triggering a string of failures in other financial institutions. The growing interconnectedness of business and financial institutions has heightened the need for measures and analytics for systemic risk measurement. The literature on techniques and metrics for assessing and managing systemic risk is nascent, and several risk measures are being proposed in this domain—see [5]. The need for systemic analysis, in addition to the analysis of individual institutions, is a growing focus of risk managers and regulators.

We define "systemic analysis" as the measurement and analysis of relationships across entities with a view to understanding the impact of these relationships on the system as a whole. The failure of a major player in a market that causes the failure/weakness of other players is an example of a systemic effect, such as that experienced with the bankruptcy of Lehman Brothers on September 15, 2008.[1]

A major challenge that makes systemic analysis harder to undertake is that it requires *most* or *all* of the data in the system—if a proper analysis of system-wide effects is to be carried out, then the data must represent the entire system. Thus, high-quality information extraction and integration that spans the entire system is critical.

Current approaches to systemic risk have used data that is easily available across the system, i.e., stock return correlations data [2, 1, 5, 15]. These papers stop short of undertaking a formal network analysis.

Midas enables enhancing the current work in finance in the following major way. By using unstructured or semi-structured public data archived by SEC and FDIC, the nature of data that is available for systemic analysis is greatly expanded. For example, in the illustrative application in this paper, we use co-lending relationships to construct networks of relationships between banks, and then use network analysis to determine which banks pose the greatest risk to the financial system. No more will researchers in finance have to only rely on the few standard (and proprietary) data sets on stock prices that are in current use.

Co-lending Systemic Risk. Using the data provided in the SEC/FDIC filings, we construct a network of connections between financial firms based on their co-investment in loans made to other corporations or financial institutions. For example, if five banks made a joint loan, we obtain all pairwise relations and score each of them to be equal to an instance of co-lending by the pair. These relationships are modeled as an undirected network with the banks as nodes, and the edges are the total count of pairwise co-lending, aggregated across all loans. These relationships may be represented in a lending adjacency matrix $\mathbf{L} \equiv \{L_{ij}\}, i, j = 1 \ldots N$, where N is the total number of financial institutions. Given that the network graph is undirected, this matrix is symmetric about its diagonal, and we set the diagonal to be zero, i.e., ignore self-loops.

We define the total lending impact on the system for each bank as $x_i, i = 1 \ldots N$. The failure of any bank i will impact the lending system by the partial withdrawal of lending support for other banks as well. Any one bank's failure will directly impact the co-lending activity of all banks it is connected with, and will also indirectly impact the banks that are connected to the ones it is directly connected with. Therefore, even if a bank has very few co-lending relationships itself, it may impact the entire system if it is connected to a few major lenders. Since the matrix $\mathbf{L}$ represents the pairwise connectedness of all banks, we may write the impact of bank i on the system as the following equa-

[1]This filing was the largest bankruptcy in the history of the U.S. financial markets.

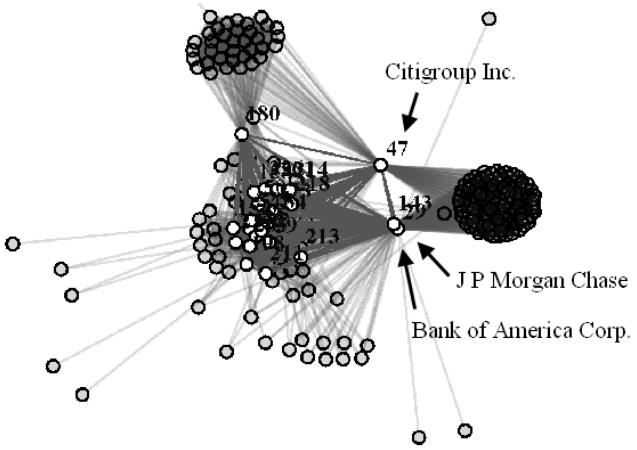


Figure 1: Co-lending network for 2005.

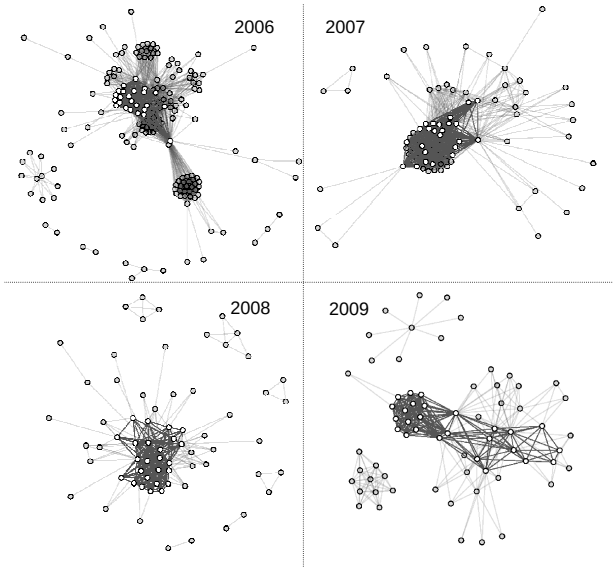


Figure 2: Co-lending networks for 2006–2009.

tion: $x_i = \sum_{j=1}^{N} L_{ij} x_j, \forall i$. This may be compactly represented as $\mathbf{x} = \mathbf{L} \cdot \mathbf{x}$, where $\mathbf{x} = [x_1, x_2, \ldots, x_N]' \in R^{N \times 1}$ and $\mathbf{L} \in R^{N \times N}$. We pre-multiply the left-hand-side of the equation above by a scalar λ to get $\lambda\ \mathbf{x} = \mathbf{L} \cdot \mathbf{x}$, i.e., an eigensystem. The principal eigenvector in this system gives the loadings of each bank on the main eigenvalue and represents the influence of each bank on the lending network. This is known as the "centrality" vector in the sociology literature [6] and delivers a measure of the systemic effect a single bank may have on the lending system. Federal regulators may use the centrality scores of all banks to rank banks in terms of their risk contribution to the entire system and determine the best allocation of supervisory attention.

The data we use comprises a sample of loans filings made by financial institutions with the SEC. Our data covers a period of five years, from 2005–2009. We look at loans between financial institutions only. Examples of included loans are 364-day bridge loans, longer term credit arrangements, Libor notes, etc. The number of loans each year is not as large as evidenced in the overnight market, and these loans are largely "co-loans", i.e., loans where several lenders jointly lend to a borrower. By examining the network of co-lenders, we may determine which ones are more critical, and we may then examine how the failure of a critical lender might damage the entire co-lending system. This offers a measure of systemic risk that is based directly on an interconnected lending mechanism, unlike indirect measures of systemic risk based on correlations of stock returns ([1]; [2]; [5]; [15]). A future extension of this analysis will look at loan amounts, whereas the current analysis is based on loan counts for which robust data is available.

After constructing the adjacency matrix representing co-lending activity, we removed all edges with weights less than 2, to eliminate banks that are minimally active in taking on lending risk with other banks. (This threshold level may be varied as required by a regulator.) We then removed all nodes that have no edges.

An example of the resulting co-lending network is presented in Figure 1 for 2005. We see that there are three large components of co-lenders, and three hub banks, with connections to the large components. There are also satellite co-lenders. In order to determine which banks in the network are most likely to contribute to systemic failure, we compute the normalized eigenvalue centrality score described previously, and report this for the top 25 banks. These are presented in Table 1. The three nodes with the highest centrality are seen to be critical hubs in the network—these are J.P. Morgan (node 143), Bank of America (node 29), and Citigroup (node 47). They are bridges between all banks, and contribute highly to systemic risk.

Figure 2 shows how the network evolves in the four years after 2005. Comparing 2006 with 2005 (Figure 1), we see that there still are disjointed large components connected by a few central nodes. From 2007 onwards, as the financial crisis begins to take hold, co-lending activity diminished markedly. Also, all high centrality banks tend to cluster into a single large giant component in the latter years.

We also compute a metric of *fragility* for the network as a whole, i.e., how quickly will the failure of any bank trigger failures across the network by expanding ripples across neighborhoods? One such metric of systemic risk is the expected degree of neighboring nodes averaged across all nodes—derived in [13], page 190, this is equal to $E(d^2)/E(d) \equiv R$, where d stands for the degree of a node. Neighborhoods are expected to expand when $R \geq 2$. We compute this for each year in our sample (Table 1). The ratio is highest just before the crisis—and then dissipates as banks take on less risk through the crisis. The diameter of the co-lending graph becomes marginally smaller as the network shrinks over time. This framework may be extended to other metrics of systemic risk to develop a systemic risk management system for regulators.

2.2 Drill-Down into Individual Entities

In this section we describe additional views that Midas provides centered around individual entities. For example, once a company such as Citigroup Inc. has been identified as a critical hub for the financial system, a regulator may want to dive deeper into various aspects that define Citigroup: its relationships with other companies (subsidiaries, competitors, investments, borrowers, etc.), its key executives (officers and directors, over the years), or aggregated financial data (loans, size of institutional investments, etc.).

For each view that we describe, we briefly mention the

Table 1: Summary statistics and the top 25 banks ordered on eigenvalue centrality for 2005.

Year	#Colending banks	#Coloans	Colending pairs	$R = E(d^2)/E(d)$	Diam.
2005	241	75	10997	137.91	5
2006	171	95	4420	172.45	5
2007	85	49	1793	73.62	4
2008	69	84	681	68.14	4
2009	69	42	598	35.35	4

(Year = 2005)

Node #	Financial Institution	Normalized Centrality
143	J P Morgan Chase & Co.	1.000
29	Bank of America Corp.	0.926
47	Citigroup Inc.	0.639
85	Deutsche Bank Ag New York Branch	0.636
225	Wachovia Bank NA	0.617
235	The Bank of New York	0.573
134	Hsbc Bank USA	0.530
39	Barclays Bank Plc	0.530
152	Keycorp	0.524
241	The Royal Bank of Scotland Plc	0.523
6	Abn Amro Bank N.V.	0.448
173	Merrill Lynch Bank USA	0.374
198	PNC Financial Services Group Inc	0.372
180	Morgan Stanley	0.362
42	Bnp Paribas	0.337
205	Royal Bank of Canada	0.289
236	The Bank of Nova Scotia	0.289
218	U.S. Bank NA	0.284
50	Calyon New York Branch	0.273
158	Lehman Brothers Bank Fsb	0.270
213	Sumitomo Mitsui Banking	0.236
214	Suntrust Banks Inc	0.232
221	UBS Loan Finance Llc	0.221
211	State Street Corp	0.210
228	Wells Fargo Bank NA	0.198

type of source documents from where the data is aggregated. The actual details and challenges regarding the various analysis stages will be described in subsequent sections.

2.2.1 Company Relationships

Figure 3 shows Citigroup's relationships with other companies through investment, lending and ownership relationships. For each relationship type, we show up to five representative companies, and also indicate the total count of related companies. The relationship types are:

- **Banking subsidiaries** : Citigroup has four banking subsidiaries registered with the FDIC. This information was obtained by integrating data from SEC and FDIC.
- **Subsidiaries** : An exhaustive list of Citigroup's global subsidiaries, as reported in their latest annual report (typically in text or html format).
- **5% Beneficial Ownership** : Securities in which Citigroup has more than 5% ownership based on analysis of SC-13D and SC-13G text filings made by Citigroup and its subsidiaries.
- **Overlapping board members/officers** : Key officer and board membership information is extracted from annual reports, proxy statements, current reports and insider transactions (text, html and xml formats).
- **Institutional Holdings** : Securities in which Citigroup has invested more than $10 million based on analysis of 13F text filings.

While the company relationship graph provides a birds-eye view of Citigroup's key relationships, additional details on individual relationships are available as described next.

2.2.2 Insider Analysis

Understanding management structure of companies and relationships across companies through common officers and board of directors is relevant in firm dynamics and corporate governance. Connected firms appear to end up merging

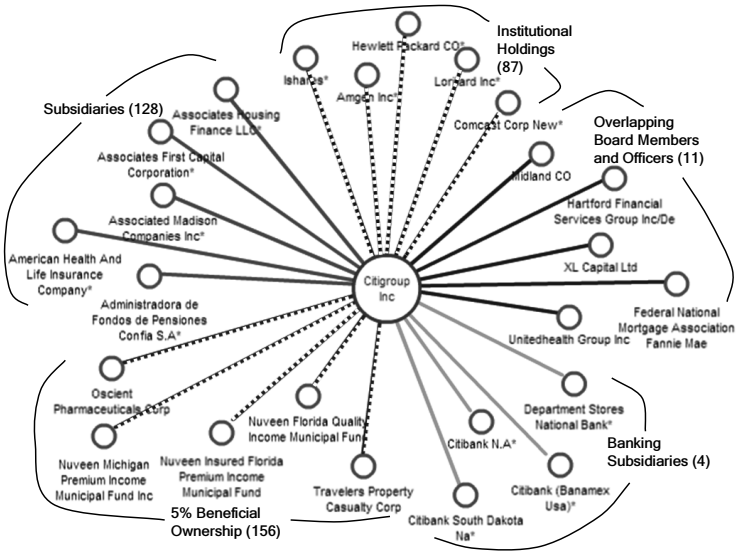


Figure 3: Companies related to Citigroup.

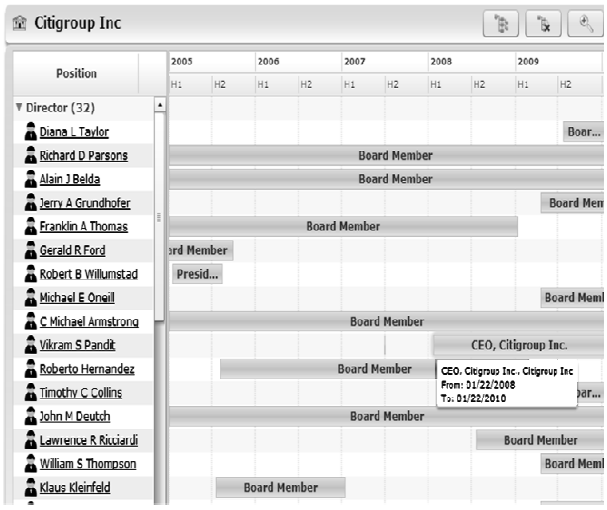


Figure 4: Key people for Citigroup.

more [7]. Understanding post-merger management structures based on earlier connections between the managers of the merged firms is also being studied [12]. To enable such analysis, Midas exposes detailed employment history and trading information for insiders (i.e., key officers and directors) of individual companies.

Employment History: Figure 4 shows some of the key officers and directors associated with Citigroup over the last several years. For each related key person, the various positions (s)he held in Citigroup along with the corresponding time periods are displayed in the figure. This profile is built by aggregating data from individual employment records present in annual reports, proxy statements, current reports and insider reports.

Insider Holdings: Figure 5 shows the current holdings of Citigroup securities (stocks and options) by the company's insiders. Each stacked bar represents the security holdings for an officer or director of Citigroup, broken down by type of holding. We show common stock, derivatives and other securities separately, with common stock further classified by whether ownership is direct or indirect (through trusts,

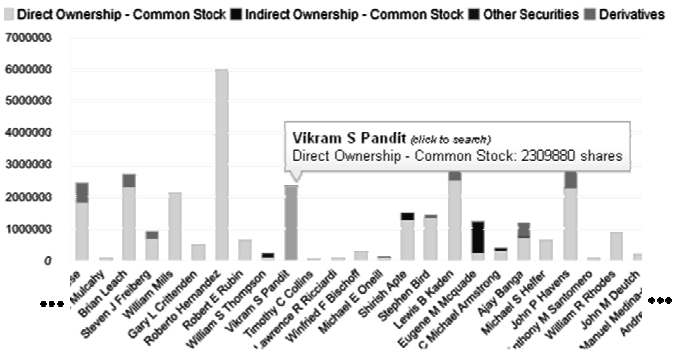


Figure 5: Insider Holdings for Citigroup.

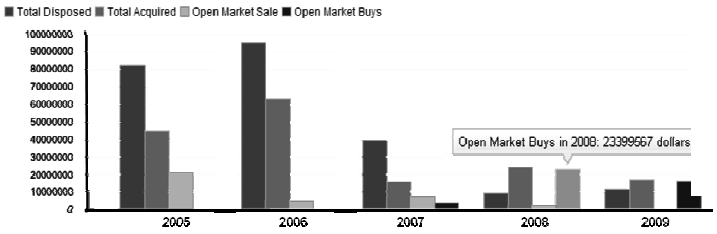


Figure 6: Insider transactions trend for Citigroup.

401K or family members).

Insider Transactions: Figure 6 presents a summary of insider transactions (buys and sells) of Citigroup securities from 2005-2009. A further breakdown of open market transactions compared with total transactions is provided. In general an open market purchase is a stronger indication of an insider's confidence. Observe that while in 2005 and 2006 there were a lot of sells of stock, in 2008 and 2009 there are not only more buys than sells, but the purchases are mostly on the open market, a very strong indication of confidence. This year so far there are more sells then buys, indicating that the trend has again reversed.

2.2.3 *Lending Exposure Analysis*

Figure 7 (top) shows a list of recent loans issued by Citigroup, either directly or through its subsidiaries. For each loan, the chart shows Citigroup's commitments to various borrowers, as compared to other co-lenders. This information has been extracted from the SEC filings made by the borrowers, where the loan documents were filed as part of their annual and current reports.

For any particular loan, additional details on the commitments made by all the lenders involved in that loan are displayed in the lower part of the figure. In this example, it shows details of an 800 million dollar loan to Charles Schwab corporation made jointly by 12 banks, including Citibank National Association, a subsidiary of Citigroup.

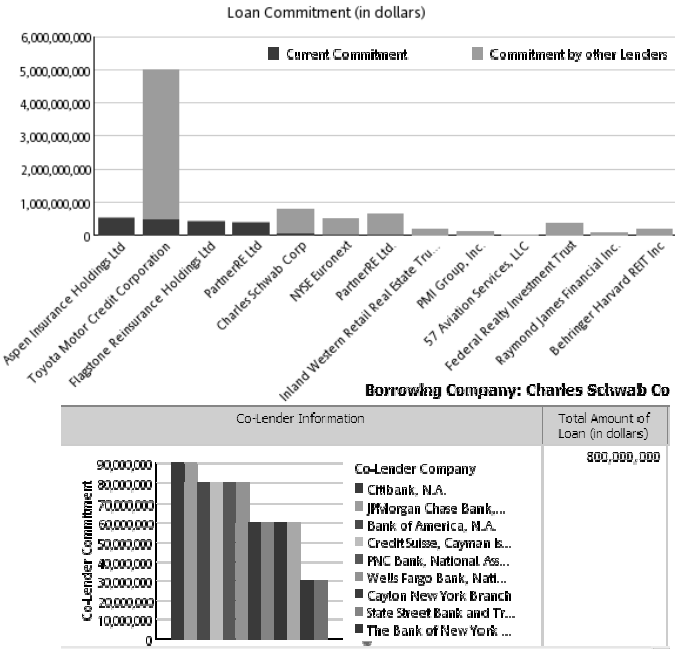


Figure 7: Lending activity for Citigroup.

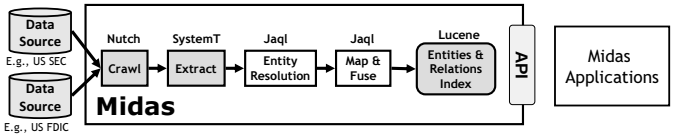


Figure 8: The Midas Data Flow

3. MIDAS OVERVIEW

We now give an overview of Midas, our system for extracting and integrating information from heterogeneous data sources. Figure 8 shows, at a high-level, the Midas data flow. Midas can take as input data from multiple sources and represented in different data formats. As output, Midas produces sets of integrated and cleansed objects and relationships between those objects which are then used by applications like the ones described in the previous section.

Input data sources can be large (Peta-bytes of information) with new incremental updates arriving daily. All operators in the Midas data flow must be capable to process large amounts of data efficiently and should scale well with increasing data sizes. To address these challenges, Midas operators are designed to run on top of Hadoop and are compiled into sequences of map/reduce jobs. For instance, the **Crawl** operator uses Nutch to retrieve input data documents. Nutch jobs are compiled into Hadoop jobs and executed in parallel. The **Extract** operator use SystemT [14] to annotate each document retrieved by **Crawl**. This operator is trivially parallelizable with Hadoop. However, the other operators (**Entity Resolution**, **Map & Fuse**) require complex data transformation whose parallel and distributed execution plan might not be trivial. To address this challenge, all instances of these operators are currently implemented using Jaql [3], a general-purpose language for data transformations. Jaql uses JSON as its data model and features a compiler that creates efficient map/reduce (Hadoop) jobs. Jaql runs the compiled jobs directly on our Hadoop cluster. Moreover, Jaql is implemented in Java and allowing many customizable extensions to be implemented in Java (e.g., user-defined functions) and seamlessly used at runtime. The Midas architecture is inspired, in part, by our Content Analytics Platform [4].

Crawl is in charge of retrieving data directly from public data sources and storing it in our local file system. Instances of **Crawl** are implemented using Nutch, a widely used open-

source crawler(http://nutch.apache.org/). To improve performance, we run Nutch as Hadoop jobs and parallelize the fetching of documents.

Extract is in charge of annotating unstructured data. Here, we leverage a large library of previously existing information extraction modules (annotators) implemented on top of SystemT [8]. SystemT is a rule-based information extraction system developed at IBM Research that makes information extraction orders of magnitude more scalable and easy to use. The system is built around AQL, a declarative rule language with a familiar SQL-like syntax. Rule developers focus on what to extract, with SystemT's cost-based optimizer determining the most efficient execution plan for the annotator. SystemT can deliver an order of magnitude higher annotation throughput compared to a state-of-the-art grammar-based IE system [8] and high-quality annotators can be built for individual domains that deliver accuracy matching or outperforming the best published results [9]. AQL rules are applied to each input document and produce a stream of annotated objects. For example, if we apply name extraction rules to the input data, we obtain structured objects that contain: 1) the raw text of the document and 2) the list of names extracted from the raw text (plus some meta-data such as the text location of each name).

Entity Resolution identifies and links annotated objects that correspond to the same real-world entity. Typically, the data required to build a single entity (e.g., a company) appears fragmented across several documents and spread over time. Recognizing that separate mentions refer to the same entity requires complex and domain-dependent analysis in which exact matching of values may not work. For instance, names of companies and people may not appear spelled the same in all documents and the documents might not explicitly contain a key to identify the company or person. Entity Resolution, which appears in the literature under other names (Record Linkage, Record Matching, Merge/Purge, De-duplication) [11], is often solved with methods that score fuzzy matches between two or more candidate records and use statistical weights to determine when these records indeed represent the same entity. Other methods explicitly use rules to express when two or more candidate records match. Our current implementation of Midas uses this latter approach and we implemented the matching rules in Jaql.

Map & Fuse transforms annotated (and possibly linked) data into a set of objects and relationships between those objects. All necessary queries to join and map the source data into the expected target schema(s) are implemented on top of this operator. The resulting queries, which are currently implemented in Jaql, must group, aggregate, and merge data into the proper, potentially nested, output schema. Since data is collected from multiple sources, duplicated values for certain fields are inevitable and must be dealt with in this stage. This data fusion step determines which of these multiple values survives and becomes the final value for the attribute. In certain cases, the data values must be merged into one consistent new value. For example, when the input set of values for a particular attribute represent time periods, we might need to compute the enclosing time period from all the valid time periods in the input set.

4. PUBLIC DATA SOURCES

Our financial application uses documents from two government data sources: the US Securities and Exchange Commission (SEC) and the US Federal Deposit Insurance Corporation (FDIC). The SEC regulates all security transactions in the US and the FDIC regulates banking institutions.

4.1 The SEC data

Public companies in the US (and key people related to these companies) are required to regularly report certain transactions with the SEC. The SEC maintains a repository of these *fillings*, organized by year and company[2]. Depending on the kind of transaction reported, public entities use different *forms* to report these regulated transactions. In some cases, forms are XML documents and, thus, contain some structured data items. In many other cases forms are filed as raw English text or as HTML documents. The SEC electronic repository contains filings going back to 1993 and currently contains over 9,000,000 filings covering about 17,000 companies and about 250,000 individual[3]. New filings are added daily and all data in the repository can be accessed via ftp.

There are many kinds of forms filed with the SEC[4] but we are only interested in those about the financial health of companies, insider transactions, and investments. We now describe the forms we used in our analysis to give a flavor of the data heterogeneity challenges we faced.

Insider Transactions (Forms 3, 4, and 5). Forms 3, 4, and 5 are XML forms that report any transaction involving securities of public company and key officer, director, or any party with at least a 10% stake on the company. These reports are filed by the company itself on behalf of the insider who is often a person but can also be another company. Form 3 is used to report when an insider is granted securities related to the company, Form 4 is used to report a transaction of such securities, and Form 5 is used annually to report all current insiders. Each form contains a common header section that provides the name of the insider, its role within the company (whether it is a key officer, directory, or a 10% owner), the name of the company, and, importantly, the *cik* for both the person and the company. The *cik* (Central Index Key) is a unique identifier provided by the SEC to every person and company that files data with the SEC. Since Forms 3/4/5 provides identifying information for both companies and key people (and due to its regulatory nature are expected to be correct), we use these forms to seed and initially populate our company and key people entities.

Financial Data and Company Status (Forms DEF 14A, 10-K, 10-Q and 8-K). Detailed information about the companies is found in a number of separate fillings. Proxy statements (Form DEF 14A) contain information for shareholders about the financial health of the company and the biographies of many key officers and directors. Much of this information is also found in the company's annual report (Form 10-K). Together, Forms 10-K and DEF 14A provide detailed business and financial information about the company including key merger and acquisitions, changes of officers and directors, business directions, key financial tables (e.g., balance sheet and income statements), executive compensation, and loan agreements. Companies must also provide quarterly updates to all shareholders, which are filed using Form 10-Q. Finally, Form 8-K is used to report signifi-

[2]http://www.sec.gov/edgar/searchedgar/webusers.htm
[3]Not all these companies or person are currently active.
[4]See http://www.sec.gov/info/edgar/forms/edgform.pdf for a complete list of all forms types.

cant events occurring in the middle of quarters. These events include mergers and acquisitions, changes of key officers or directors, offerings of equity/debt, bankruptcy, and entering material definitive agreements. All these forms contain a header that identify the company filing the form (including its *cik*). The content of the report is, however, English text formatted with HTML. Some of the financial tables are now reported in XBRL (XML, see http://xbrl.org/), but this is a recent requirement and many legacy filings in the repository contain this data in HTML tables.

Institutional Investment (Forms 13F, SC 13D and SC 13G). Companies report quarterly their ownership of securities in other companies. Form 13F, the institutional investment report, states each security owed by the reporting company, including the number of shares and the kind of share, in fixed-length column table format. However, the table representation varies from filer to filer making the task of identifying the columns and values a challenge. Form SC 13D and SC 13G are used to report 5% owners of securities related to the filing company.

4.2 The FDIC data

US banking institutions are required to report their financial health to the FDIC on a quarterly basis. These reports are very structured and are filed in XBRL format. In many cases, banks are subsidiaries of the public *holding company* which reports with the SEC. That is, often the parent company of a bank reports its results with the SEC while at the same time detailed information about the bank is submitted separately with the FDIC. All data in the FDIC repository can be accessed using a published web-service[5].

5. MIDAS INTEGRATION FLOW

We now give concrete details of the Midas flow that integrates information related to financial companies. We start by discussing the process that crawls all the forms related to the financial companies. We then discuss in Section 5.2 the initial construction of a reference or core set of company and people entities from insider reports (Forms 3/4/5). Since these forms are in XML and contain structured and relatively clean data, the resulting core set of entities forms the backbone of the rest of the integration flow. In Section 5.3, we detail how further information from a myriad of unstructured forms is extracted, linked and fused to the core set of entities. The final result is a set of entities with rich relationships, including detailed employment histories of key people, lending/co-lending relationships among companies, and all the other relationships we discussed in Section 2.2.

5.1 Crawling Data

The SEC contains data about *all* public companies in the US filing since 1993. We, however, are only interested in "financial" companies. Further, to avoid having too many stale entities in our data set, we restrict our crawl to documents no more than five years old (i.e., 2005-2010). Fortunately, the SEC publishes an index of all filings in the repository that we use to decide if a document is relevant. This index, which is updated daily, contains the *cik* and name of the filing company, the type of form filed (3/4/5, 10-K, etc.), and the ftp url to the actual document.

To determine if a company is a financial company, we preprocessed a large number of 10-K reports for *all* companies filing with the SEC for a period of 2 years. On each 10-K form, companies report their "Standard Industrial Classification (SIC) Code", an industry-wide numeric classification code. Roughly, entities reporting an SIC code in the [6000-6999] range are considered financial companies[6]. Using the SIC codes, extracted a "master" list of 3,366 financial companies ciks.

Given this master cik list, a range of dates (2005-2010), and a list form type we want, we filter the daily SEC document index and identify the ftp urls we need. The list of ftp urls forms a "seed" list that is fed into Nutch for crawling. In contrast to traditional web-crawling, our target documents do not change over time. The filings are never replaced with new updated versions and, thus, Nutch does not need to revisit previously crawled pages. Moreover, the seed list contains all the documents we want to crawl and Nutch does not need to parse the crawled documents to find more links.

Crawling data from the FDIC does not require filtering by industry code since, by definition, all banks are financial institutions. The FDIC publishes a web-service that allows downloading of the current financial report of a particular bank. Our crawler is in a web-service client that regularly downloads the most recent reports for all active banks.

We currently have a repository with close to 1,000,000 SEC documents related to financial companies and 77,000 FDIC reports for active banks. The SEC imposes some limits on crawlers (e.g., we could only run the crawler overnight) and it took several months to bootstrap the system with data covering several years. We now run the SEC and FDIC crawler monthly to catch up with recent filings.

5.2 Constructing Core Entities

We now discuss the initial construction and aggregation of company and key people entities from the XML files that correspond to insider reports (Forms 3/4/5).

Extraction of records from XML forms. We use Jaql to extract (and convert to JSON) the relevant facts from XML Forms 3/4/5. Each of these facts states the relationship, as of a given reporting date, between a company and a key officer or director. The relevant attributes for the company are: the SEC key (or cik) of the company, the company name and address, the company stock symbol. The relevant attributes for the person are: the SEC key or cik, name, an attribute identifying whether the person is an officer or a director, and the title of the person (i.e., "CEO", "Executive VP", "CFO", etc) if an officer. Other important attributes include the reporting date, a document id, a list of transactions (e.g., stock buys or sells, exercise of options) that the person has executed in the reporting period, and a list of current holdings that the person has with the company.

Aggregation of company and people entities. In this step, we process all the facts that were extracted from XML forms and group them by company cik. Each group forms the skeleton for a company entity. The important attributes and relationships for a company are aggregated from the group of records with the given company cik. As an example of important attributes of a company, we aggregate the set of all officers of a company such as Citigroup Inc. This aggregation is with respect to all the forms 3/4/5 that Citigroup Inc. has filed over the five years. Additional fusion

[5]See https://cdr.ffiec.gov/public/.

[6]See http://www.sec.gov/info/edgar/siccodes.htm.

must be done so that each officer appears only once in the list. Furthermore, for each officer, we aggregate all the positions that the respective person has held with the company. As an example, a person such as Sallie Krawcheck will result in one occurrence within the list of officers of Citigroup, where this occurrence contains the list of all the positions held by Sallie Krawcheck with Citigroup (e.g., CFO, CEO of Global Wealth Management). Since positions are strings that vary across forms, normalization code is used to identify and fuse the "same" position. Finally, each position is associated with a set of dates, corresponding to all the filings that report that position. The earliest and the latest date in this set of dates is used to define the time span of the position (assuming continuous employment). The end result of this analysis is exemplified in Figure 4.

To give a quantitative feel for the above processing, there are about 400,000 facts that are aggregated. Roughly, this number corresponds to the number of forms 3/4/5 that were filed over the five-year period by all the financial companies. These 400,000 facts result in about 2,500 company entities, each with a rich structure containing officers with their position timelines (within the company), directors (with similar timelines), and also containing an aggregation of transactions and holdings (to be discussed shortly).

A separate but similar processing generates, from the same 400,000 facts, an inverted view where people are the top-level entities. We generate about 32,000 people entities, corresponding to the officers or directors that have worked for the 2,500 financial companies.Like a company, each person entity is also a complex object with nested attributes such as employment history, which spans, in general, multiple companies. For example, a person such as Sallie Krawcheck will have an employment history spanning both Citigroup Inc. (where she served as CFO and then CEO of Global Wealth Management) and Bank of America (which she joined later as President of Global Wealth and Investment Banking).

Fusion of insider transactions and holdings. The aggregation of the transaction and holding data over the collection of forms 3/4/5 requires a detailed temporal and numerical analysis. First, we need to ensure that we group together securities of the same type. In general, there are multiple types of securities (derivatives or non derivatives), types of ownership (direct or indirect), and types of transactions (acquired, disposed, granted, open market purchase, etc.). The various values for such types are reported in text and have variations (e.g., "Common Stock" vs. "Class A stock" vs. "Common shares"). In order to avoid double counting of transactions and to report only the most recent holding amount for each type, we developed normalization code for types of securities and for types of ownership. Subsequent processing summarizes, for each company entity and for each year, the total amount of transactions of certain type (e.g., open market purchases) that company insiders executed in that year. The results of such aggregation were shown earlier in Figure 6. Similar processing retains, for each person entity, the current (i.e., the most recent) holding that the person has with a given company, for each type of securities (Figure 5).

5.3 Incorporating Data from Unstructured Forms

We now discuss the processing involved in the extraction and fusion of new facts from unstructured data into the core entities. The new facts, which are extracted from either

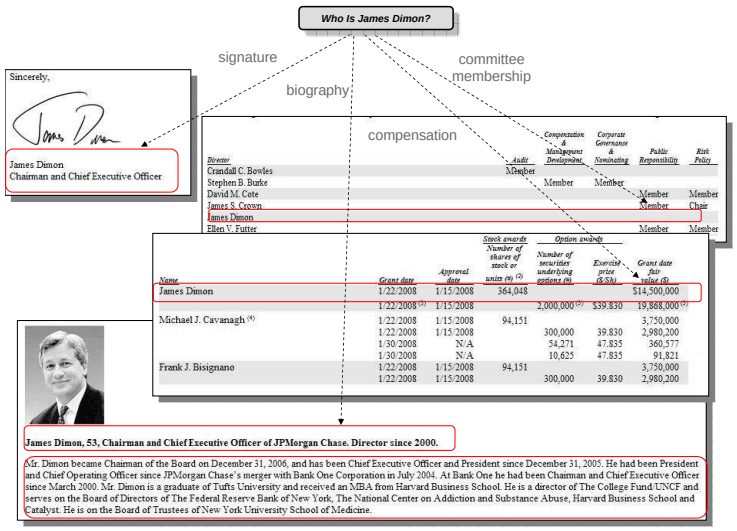


Figure 9: Employment information in various filings

text or tables, describe new attributes or relationships, and typically mention a company or a person by name without, necessarily, a key. Thus, before the new information can be fused into the existing data, entity resolution is needed to perform the linkage from the entity mentions to the actual entities in the core set.

5.3.1 *Example 1 : Enriching Employment History*

In addition to the insider reports, information about a person's association with a company is present in a wide variety of less structured filings, as illustrated in Figure 9. This information ranges from point-in-time facts (when an officer/director signs a document) to complete biographies that provide the employment history of a person. To extract and correctly fuse all the needed pieces of information, we must address several challenges.

Extract. Employment history records need to be extracted from various contexts such as biographies, signatures, job change announcements, and committee membership and compensation data. These records are typically of the form (person name, position, company name, start date, end date) for each position mentioned in the text. However, not all of the attribute values may be present or extracted successfully. For instance, the expected output from the biography in Figure 9 would include (James Dimon, Chairman, JP Morgan Chase, –, –), (James Dimon, Chief Executive Officer, JP Morgan Chase, –, –), (James Dimon, Director, JP Morgan Chase, 2000, –) and (Mr. Dimon, Chairman, unknown, "December 31, 2006", –). Using biographies as an example, we illustrate some of the challenges we encounter in extracting employment records from unstructured documents.

Identifying the beginning of a biography. Biographies typically appear in annual reports and proxy statements, as short paragraphs within very large HTML documents (100's KBs to 10s MBs) and within HTML tables, where individual employment facts may be formatted in different ways. For instance, a position with a long title may span multiple rows while the corresponding person's name may align with one of these rows, depending on the desired visual layout.

Past positions are expressed differently. For instance, a set of positions may be linked with a single organization (Chair-

man and Chief Executive Officer of JP Morgan Chase) or multiple positions may be associated with a single start date (Chief Executive Officer and President since 12/31/2005).

Anaphora resolution. Individual sentences may refer to an individual via a partial name (e.g., "Mr. Dimon") or by using pronouns (e.g., "he"). Sometime the name of a related individual may be mentioned in the biography.

Based on 10 random samples of all DEF 14A filings, our biographies annotator obtains 87% precision and 49% recall for extracting key people's names, and 91% precision and 51% recall for extracting the correct block of biographies.

Entity Resolution. As mentioned, the attributes extracted for biographies include the name of the person, the name of the filer company (also the cik, since this is associated with the filing entity) and the biography text itself. However, information in biographies does not contain a cik for the person and we need entity resolution to link each extracted biography record to a person cik.

Entity resolution is an iterative process requiring a complex and domain-dependent analysis that requires understanding the data, writing and tuning entity resolution rules, and evaluating the resulting precision (are all matches correct?) and recall (did we miss any matches and why?). In the process of matching people mentioned in biographies to the actual people entities, we faced the following challenges:

No standardization in entity names. People names come in different formats (e.g. "John A. Thain" vs. "Thain John" vs. "Mr. Thain", or "Murphy David J" vs. "Murphy David James III"). Hence, exact name matching will only find some matches and we need approximate name matching functions to resolve more biographies. On the other hand, two people with similar names (even when working for the same company) may be in fact two different people. For example, "Murphy David J" and "Murphy David James III" are two different people. *To tackle this challenge*, we designed specialized person name normalization and matching functions that cater for variations in names, suffixes such as "Jr.", 'II", and allow matching names at varying precision levels. We iterated through our data and entity resolution results several times in order to fine-tune our functions.

Achieving high precision. To improve precision beyond just the use of name matching, we observed that for a biography record, we typically know the cik of the company (since it is the filing entity). As a result, we were able to develop matching rules that exploit such contextual information. In particular, the rules narrow the scope of matching to only consider the people entities that are already known to be officers or directors of the filing company (as computed from Forms 3/4/5).

Improving recall. To improve recall, in general, one needs multiple entity resolution rules. For example, there are cases where the filer company is not in the employment history of a person (based on Forms 3/4/5). To account for such case, we had to include other, more relaxed rules that were based just on name matching. Having multiple rules, we prioritized them so that weaker matches are kept only when we do not have any matches based on stronger evidence. For instance, if we matched a "Thain John A" mentioned in a biography to both a "John A. Thain" and a "Thain John" in key people, via two different rules, we will only keep the first match since it is based on a rule that matches first/lastname *and* middlename initial.

Our initial matching rules achieved a 82.29% recall, that is, 82.29% of 23,195 biographies were matched to a person cik. At the end of the tuning process, we raised that to 97.38%. We measured precision by sampling our data, and we found it to be close to 100%.

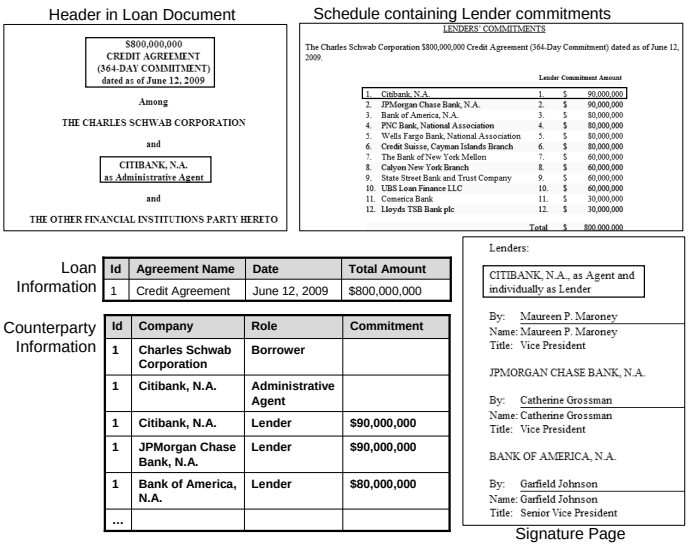


Figure 10: Loan document and extracted data

5.3.2 *Example 2 : Lending Exposure Analysis*

Figure 10 shows portions of a loan document filed by Charles Schwab Corporation with the SEC. This loan document is a complex 70 page HTML document, that contains key information about the loan such as the loan amount, date the agreement was signed, the companies involved in various capacities and the commitment made by individual lenders. As shown in the figure, this information is spread across different portions of the document such as the header at the beginning of the loan document, signature page and schedules containing lender commitments. The following analysis steps are performed on the loan data.

Extract. We first identify documents that describe loan agreements. Additional rules extract basic loan information from the header portion of these documents, which may appear either in a tabular form (as shown in this example) or as a paragraph in free-flowing text. The names and roles of the various counterparties involved in the loan are identified from three portions of the loan — header, signature and commitment table. Finally, the dollar amounts committed by individual lenders are extracted from commitment tables that typically appear in html tables in the document. Additional details about the name and role of officers who signed the loan document on behalf of different companies are also extracted. Portions of the extracted data for loan and counterparty information are shown in the figure.

Entity Resolution. Each extracted fact contains one or more company and person names, whose real-world identity needs to be resolved to facilitate aggregating facts from all loan documents. For example, for identifying lenders, we faced the following challenges.

Company name variations and subsidiaries. Company names may be written in various forms, for example, "Citibank, N.A.", and "CitiBank National Association". In addition, companies have subsidiaries; for example both "Citigroup Global Markets, Inc" and "CitiBank National Association"

are subsidiaries of Citigroup Inc. We need to be able to say when two company names refer to the same company and when one is a subsidiary of the other. To determine the unique identity of each lender, we built special normalization functions for company names and rules that compare the names of lenders with the names of all companies filing with the SEC and FDIC, and the names of all of their subsidiaries (extracted from the annual reports).

Measuring recall is another challenge because 1) we could indeed fail to resolve a company that is a lender, or 2) a company mentioned in a loan document does not file with SEC or it is not a lender. Unfortunately, in the latter case, we do not have the role of each company we extract from loan documents. We sampled 60 companies from our list of companies extracted from loan documents; 17% of them were resolved and they were all correct (i.e., achieving 100% precision); and 12.69% were not resolved but these contained errors from information extraction. Hence, our entity resolution rules are robust and do not propagate errors generated in the previous phase. 26.9% were companies that do not file with SEC hence, we do not resolve them. Finally, 42.8% were not resolved and included companies that are borrowers or institutions with no lending capacity.

Constructing the co-lending matrix. Based on the information extracted from loan documents, we were able to construct, for each year, a co-lending network where the nodes are lenders and an edge between two nodes counts the total number of loans where the two entities are co-lenders. One of the challenges in building a meaningful network is to generate a single node per company, since in the source data, a lender can appear under multiple names. For example, "Citibank" and "Citicorp USA" must be fused into the same entity ("Citigroup Inc.", which is the parent company). Entity resolution enables us to perform such identification. Once the nodes are correctly fused, subsequent processing computes the aggregated count of loans for each pair of nodes. The resulting co-lending matrix forms the basis for the systemic risk analysis described in Section 2.1.

6. OTHER BUSINESS APPLICATIONS

We conclude this paper with a description of some business applications that can exploit the consolidated public data from Midas, enhanced with more unstructured public data such as blogs, message boards, news feeds, etc.

Risk Measurement: In Section 2 we showed that financial institution systemic risk metrics may be developed from an analysis of the network of bank co-lending relationships. Measures such as centrality will help identify banks that are critical in the lending system. Community detection in lender networks will uncover groups of lenders that are critical to the system.

Generating non-return based data: Most public data is not available in structured data sets, nor is it widely available in numerical form. Text discussion on message boards, blogs, news forums, etc., can be used to uncover connectedness between firms and banks. Hence, construction of new data sets for meeting analysis or regulatory goals is an important application. Midas has already demonstrated several use cases in this domain.

Analyzing Organization Structure: Relationships between CEOs and management officers of firms are now being shown to be relevant in firm dynamics and corporate governance. Connected firms appear to end up merging more [7]. Understanding post-merger management structures based on earlier connections between the managers of the merged firms is also being studied [12].

Supporting Regulation: Large-scale data integration for decision-making and regulation is a growing field. In finance, the establishment of the National Institute of Finance (NIF) under the auspices of the Office of Financial Research (OFR), proposed in the Restoring American Financial Stability Act[7], has been tasked with setting up a systemic risk data warehouse for just this purpose. Technologies such as Midas are therefore extremely timely and may be deployed by the OFR.

Trading: Developing statistical arbitrage signals for convergence trading and high-frequency trading. This will be based on extracting signals from news feeds, blogs, message boards, and other public opinion forums.

7. REFERENCES

[1] V. Acharya, L. Pedersen, T. Philippon, and M. Richardson. Measuring Systemic Risk. SSRN: http://ssrn.com/abstract=1573171, 2010.

[2] T. Adrian and M. Brunnermeier. CoVaR. http://www.princeton.edu/∼markus/research/papers/CoVaR.pdf, Princeton University, 2009.

[3] K. Beyer and V. Ercegovac. Jaql: A Query Language for JSON, 2009. http://code.google.com/p/jaql/.

[4] K. S. Beyer, V. Ercegovac, R. Krishnamurthy, S. Raghavan, J. Rao, F. Reiss, E. J. Shekita, D. E. Simmen, S. Tata, S. Vaithyanathan, and H. Zhu. Towards a Scalable Enterprise Content Analytics Platform. *IEEE Data Eng. Bull.*, 32(1):28–35, 2009.

[5] M. Billio, M. Getmansky, A. Lo, and L. Pelizzon. Econometric Measures of Systemic Risk in the Finance and Insurance Sectors. SSRN: http://ssrn.com/abstract=1648023, 2010.

[6] P. Bonacich. Power and centrality: A family of measures. *American Journal of Sociology*, 92(5):1170–1182, 1987.

[7] Y. Cai and M. Sevilir. Board Connections and M&A Transactions. SSRN: http://ssrn.com/abstract=1491064, 2009.

[8] L. Chiticariu, R. Krishnamurthy, Y. Li, S. Raghavan, F. Reiss, and S. Vaithyanathan. SystemT: An Algebraic Approach to Declarative Information Extraction. In *ACL*, 2010.

[9] L. Chiticariu, R. Krishnamurthy, Y. Li, F. Reiss, and S. Vaithyanathan. Domain Adaptation of Rule-Based Annotators for Named-Entity Recognition Tasks. In *EMNLP*, 2010.

[10] N. N. Dalvi, R. Kumar, B. Pang, R. Ramakrishnan, A. Tomkins, P. Bohannon, S. Keerthi, and S. Merugu. A Web of Concepts. In *PODS*, pages 1–12, 2009.

[11] A. K. Elmagarmid, P. G. Ipeirotis, and V. S. Verykios. Duplicate Record Detection: A Survey. *IEEE Trans. Knowl. Data Eng.*, 19(1):1–16, 2007.

[12] B.-H. Hwang and S. Kim. It pays to have friends. *Journal of Financial Economics*, 93(1):138–158, 2009.

[13] M. Jackson. *Social and Economic Networks*. Princeton University Press, NJ, 2009.

[14] R. Krishnamurthy, Y. Li, S. Raghavan, F. Reiss, S. Vaithyanathan, and H. Zhu. SystemT: A System for Declarative Information Extraction. *SIGMOD Record*, 37(4):7–13, 2008.

[15] M. Kritzman, Y. Li, S. Page, and R. Rigobon. Principal components as a measure of systemic risk. SSRN: http://ssrn.com/abstract=1582687, 2010.

[7]See `http://www.ce-nif.org/`