Visual Analysis of Venture Similarity in Entrepreneurial Ecosystems

Rahul C. Basole, Senior Member, IEEE, Hyunwoo Park, and Raul O. Chao

Abstract—Entrepreneurial ecosystems are vital sources of innovation and critical engines for economic growth. In this study, we use text-based analysis, network visualizations, and topic modeling of nearly 60,000 venture business descriptions—i.e., how ventures present technologies and products to key stakeholders including customers, employees, and investors—to examine the structure of 35 global entrepreneurial ecosystems. Rather than using predefined industry classifications, we allow the structural configurations to emerge endogenously revealing a more variegated perspective of venture strategic positioning in entrepreneurial ecosystems. Our study makes several important contributions. First, by examining strategic positioning statements of geographically defined ventures, we contribute and advance our understanding of the geography of innovation and structure of entrepreneurial ecosystems. Our results indicate that there are wide differences in entrepreneurial ecosystem size, structure, composition, and venture strategic positioning. Second, methodologically, we use novel computational approaches and introduce visualization as a powerful means to understand entrepreneurial ecosystems. Third, our results show that ventures from widely different industries often use similar position statements, thus highlighting that ecosystems are indeed not just defined by industries, but also strategic positioning. We conclude with theoretical and managerial implications.

Index Terms—Cluster analysis, entrepreneurial ecosystems, strategic positioning, text analytics, visualization.

I. INTRODUCTION

ENTREPRENEURIAL ecosystems have been a topic of great interest to both innovation and technology management scholars, as well as entrepreneurs, policymakers, and venture capitalists [1]–[4]. This interest is in part fueled by the fact that entrepreneurial ecosystems are not just vital sources of innovation but also critical engines for economic growth [5]. Arguably, the example of the most prominent entrepreneurial ecosystem is Silicon Valley. It is often touted to be the “gold standard” ecosystem, creating many game-changing giants such as Intel, Google, Apple, and Facebook [6] and inspiring the emergence of many other ecosystems across the globe [7].

An entrepreneurial ecosystem can be defined as a set of interdependent organizations that engage in productive entrepreneurship within a geographically defined territory [8], [9]. Productive entrepreneurship refers to the outcome of ambitious entrepreneurs who pursue opportunities to create value-added products and services [10]. The ecosystem metaphor stresses that this value creation is generated not by a single organization but rather within a network of symbiotically interconnected organizations [11]. The spatial context suggests that these entrepreneurial activities typically occur in close geographic proximity, creating valuable agglomeration and network spillover effects [12].

Our understanding of entrepreneurial ecosystems is rooted in two well-established streams of research: 1) the regional development literature; and 2) the strategy literature [13]. The regional development literature has focused on explaining the differential socioeconomic performance of regions, examining concepts like industrial districts, regional industrial clusters, agglomerations, and regional systems of innovation [5], [14], [15]. The strategy literature, on the other hand, has focused on business ecosystems as a form of economic coordination in which firms collaborate and compete with each other to create an appropriate value [11], [16], [17]. While much has been learned about entrepreneurial ecosystems, research to date has predominantly focused on documenting the presence (or absence) of particular ecosystem components, such as access to venture capital funding, proximity to educational institutions, or agglomeration of industries [18], [19]. Existing work has not yet fully exploited insights into the structural nature of entrepreneurial ecosystems and the strategic positioning ventures assume [20], potentially missing important nuances of value configurations that may exist. Prior studies have indeed shown that entrepreneurial ecosystems are actually composed of a diverse set of industries, each providing unique complementary and competing value offerings [21], [22]. However, an industry-centric lens simplifies the characterization of entrepreneurial ecosystems greatly, often leading to geographic/industry stereotyping—such as biotech in Boston, fintech in London, or mobile in Singapore [23], [24].

In this study, we aim to extend our understanding of entrepreneurial ecosystems beyond this industry-centric lens by providing a more granular, structural view of ecosystem value configuration using a visual analytic approach. In contrast to
the prior work, we allow structural configurations to emerge endogenously from the business descriptions ventures use to describe their strategic positioning. The clusters that we uncover, thus, do not map to existing industry classifications, but provide a more nuanced perspective of entrepreneurial ecosystem activities. By examining the topological structure, we address the recent call to examine the relational organization of entrepreneurial ecosystems [25] and consider structure as an important construct of ecosystems [20]. We pursue this research objective by employing novel computational techniques to analyze publicly available unstructured data of nearly 60,000 ventures in 35 distinct global entrepreneurship ecosystems. Specifically, we derive ecosystem relationships between entrepreneurial ventures based on the similarity of textual content. Rather than just using a network analytic approach, we use a complementary visual lens to graphically depict the structure of entrepreneurial ecosystems, enabling us to detect structural patterns, clusters, and outliers [26]. In doing so, our visual analysis allows us to draw conclusions regarding differences between ventures in an ecosystem as well as differences between the ecosystems themselves.

Our analysis indicates that the heterogeneity of venture positions within an ecosystem varies considerably depending on the entrepreneurial ecosystem characteristics. In fact, for ventures that exist in larger ecosystems, similarity in business descriptions converges to a global mean. Additionally, we find that entrepreneurial ecosystems located in emerging economies tend to be smaller and they tend to have greater dispersion of similarity in terms of business descriptions. In other words, emerging entrepreneurial ecosystems tend to have ventures that position themselves either very similarly or very differentiated. The key implication of our study is that ventures are constantly balancing legitimacy and differentiation and that this balance becomes more salient as an entrepreneurial ecosystem grows in size.

Our study makes several contributions to the technology management, entrepreneurship, and strategy literature. First, we provide a robust and generalizable technique to describe global entrepreneurial ecosystems with a focus on both traditional measures such as size and industry concentration, as well as more novel measures that capture strategic position and organizational identity [27], [28] relative to other ventures. Positioning and organizational identity are critical as they embody how a venture identifies itself to stakeholders including customers, employees, and investors. Second, by examining strategic positioning statements of geographically defined ventures, we contribute and advance our understanding of the geography of innovation, structure of entrepreneurial ecosystems, and widely debated tradeoff between conformity and differentiation [29], [30]. Third, methodologically we use novel computational approaches and introduce visualization as a powerful means to understand entrepreneurial ecosystems. In doing so, we address the call of bringing an important new methodology to the field of innovation and entrepreneurship [31]. Our approach can be easily applied to other business ecosystem contexts, and thus, serves as an important foundation for subsequent research. Finally, our results show that ventures from widely different industries actually often use similar business descriptions, thus, highlighting that entrepreneurial ecosystems are indeed not just defined by industries but also strategic positioning.

The remainder of the paper is structured as follows. We first provide a brief review of related work on entrepreneurial ecosystems and strategic positioning. Next, we describe our methodology, including our extensive data extraction and curation process, our data mining and analysis approach, and visualization of our global entrepreneurial ecosystems. We then present and discuss our results. We conclude with implications, limitations, and future research.

II. RELATED WORK

Our study of venture similarity in entrepreneurial ecosystems demands a review of related work on industrial clusters, regional innovation systems, and business ecosystems as well as strategic positioning of ventures.

A. From Clusters and Regional Innovation Systems to Entrepreneurial Ecosystems

Understanding the structure and dynamics of spatial agglomeration of industrial and economic activities has been a topic of great interest to scholars for many decades. Commencing with Marshall’s pioneering analysis of industrial concentrations in Victorian England leading to “agglomeration economies” [32], subsequent studies have found that firms across many advantages from spatial co-location with similar firms, in particular the development of specialized pools of human capital, specialist suppliers, and specialist infrastructure. The Marshallian view is often contrasted with the Jacobian externality perspective [33], which argued that spatial agglomerations of unrelated industries can lead to knowledge spillovers, in which ideas from one industry can be applied in another. However, there continues to be inconclusive evidence whether the Marshallian specialization or Jacobian diversification most favors regional innovation [34]. Porter’s seminal work on geographical clusters [35] argued that firms benefited from both ideas, namely local sectoral specialization and knowledge spillovers. However, Saxenian’s [36] groundbreaking study of Route 128 and Silicon Valley showed that different clusters in fact operate in fundamentally different ways, underlining that potentially distinct structure and dynamics are at play. Advocated by Porter and colleagues, the “cluster” idea continued to gain traction in the regional development and economic geography literature, often used as the “holy grail” policy concept [37]–[39].

In parallel, innovation scholars developed the concept of innovation systems to understand the systemic processes that characterize localized knowledge creation and transfer [40], [41]. A key focus of this approach is its emphasis on the relational characteristics between different stakeholders and how it influences the innovation process [42], [43]. The applicability of this lens led to a significant growth in economic geography studies examining a variety of regional innovation systems.

The concept of entrepreneurial ecosystems is a direct result of the prior work on geographical clusters and regional innovation systems. Similar to the other two fields, firms play a central role in entrepreneurial ecosystems. Drawing on the bi-
ological metaphor [16], entrepreneurial ecosystems are shaped by complex interactions and interdependencies between entities and are constantly evolving. While introduced by Moore and predominantly used in the strategy domain, the concept of entrepreneurial ecosystems was popularized in the entrepreneurship literature by the pioneering work by Isenberg [44]. Similar to the study of geographical clusters and innovation systems, an entrepreneurial ecosystem adapts the idea of spatial boundedness of economic activities. However, in contrast to the prior work, the emphasis in existing entrepreneurial ecosystems is increasingly on compositional and relational aspects. Despite this interest, the literature yet lacks a comprehensive analysis of the structural underpinnings of entrepreneurial ecosystems [25].

B. Strategic Positioning

In order to succeed, ventures must develop appropriately positioned business models. A successful model captures how a venture will make money and sustain profits, how it organizes itself, what core value propositions it provides, and how it will align itself in the market relative to other stakeholders [45]. Strategic positioning is, thus, a central idea in the business model construct.

The literature on strategic positioning is primarily rooted in two domains: strategic management and marketing. The majority of studies focus on developing a theoretical framework or performing cluster analysis to describe the strategic positioning options available to firms. Almost without exception the studies define strategic positions a priori, forcing an exogenous structure upon the companies being analyzed. An inherent problem with this approach is that conventional wisdom of the domains themselves (i.e., strategy and marketing) tend to dominate the structures that emerge based on the view of the researcher. In addition, the few empirical studies that exist employ relatively small-scale proprietary datasets or individual case studies. This practice limits the possibility of broader empirical examination, both on a large scale within an ecosystem, as well as between ecosystems themselves.

Early work on strategic positioning began with the work of Miles et al. [46], who studied alternative ways that organizations define their product/market domains. At the same time, Mintzberg [47] put forth the concept of “intended strategies,” which can be interpreted as positioning statements in that they are plans made in advance of specific decisions. Porter [48] was the first to define clear strategic positioning choices when he discussed firm strategies as either narrow or broad in scope and low-cost or differentiated in terms of core capability. The result was strategic positioning based on three possible choices: overall cost leadership, cost focus, and differentiation. Markides [49], [50] adopted a slightly different perspective in terms of strategic positioning defining answers to questions about “who, what, how” the firm operates. All of the studies mentioned above are frameworks created based on theory or a limited set of individual case analyses. Deephouse [51] was among the first to develop empirical tests of strategic positioning. He studied the balance between differentiation, which reduces competition versus conformity, which provides legitimacy. Our study generalizes and expands this work as we develop a “similarity metric” between firms based on their strategic position.

As a consequence of the early work in strategy, marketing scholars took up the issue of competitive position with a particular focus on market positioning strategies. Hooley et al. [52], described how competitive positioning links the internal capabilities of the firm to external market segments. Blankson and Kalafatis [53], [54] created a typology of marketing positioning strategies. Similar to the early work in strategy, these papers are frameworks that presuppose a specific set of strategic positions. Later work in marketing conducted empirical analysis to study positioning. Kalafatis et al. [55], analyzed the elements that make up a strategic marketing position. Hooley and Greenley [56] analyzed when positioning can deliver a sustained competitive advantage. Both of these articles (and others that followed) used cluster analysis with primary survey data. The problem with this approach is that it forces structure upon the firms in question, rather than allowing that structure to emerge endogenously.

There are two important limitations with the existing work on strategic positioning. The first is that the frameworks or clusters that define groups are chosen a priori by the researcher. This necessarily forces an exogenous structure on the data. Second, if and when empirical analysis is brought to bear, it is done with relatively small proprietary datasets or individual case studies. This limits the generalization and breadth of the empirical results. Our computational approach solves both of these problems as we do not presuppose or force any structure on the data; rather we allow groups and strategic positioning statements to emerge endogenously based on a computational analysis of the similarity between firms in an ecosystem. In addition, we do this with a dataset of nearly 60,000 ventures across 35 ecosystems around the world. The scale of this analysis provides rich opportunity for understanding strategic positioning within an ecosystem as well as, and perhaps more importantly, across ecosystems.

III. METHODOLOGY

Our study uses a three-phase approach (see Fig. 1) for analyzing and visualizing strategic positioning of ventures in global entrepreneurial ecosystems, consisting of data extraction and curation, data mining and analysis, and visualization. Our approach builds on the well-established data-to-knowledge “human-in-the-loop” model by carefully balancing data management, visual encodings, and sensemaking [57]. We elaborate on each of these steps, as well as potential alternatives, in the sections that follow.

A. Data Extraction and Curation

Our study uses Crunchbase, a wiki-style curated open source directory of more than 100,000 global technology companies, people, and investors, as the primary dataset. While other data sources exist (e.g., Thomson VentureXpert, Owler, AngelList, CB Insights), Crunchbase arguably provides the most detailed, up-to-date, and open data on companies, including company description, founding year, names of executive teams, funding

1http://www.crunchbase.com
rounds and amounts, office locations (city, state) and geographic regions, industry segments, and number of employees. Crunchbase is a community-driven dataset; contributors include executives, entrepreneurs, and investors, who all actively contribute to company profile pages. While data curation is socially driven, the quality and coverage of the data is monitored and updated continuously through several means. First, it is edited and managed by executives and investors associated with the venture. Second, machine learning algorithms are used to compare data for accuracy and anomalies against other publicly-available information (including corporate websites, analyst reports, and TechCrunch). Lastly, Crunchbase employs a team of global experts and data analysts, who provide manual data validation and curation. We used the Crunchbase application programming interface (API) to extract a complete data dump of 58 880 companies worldwide. It is important to note a few data nuances. First, the data includes a geographic region field, which captures the broader geographic area rather than traditional city boundaries. For example, San Francisco, San Jose, and Palo Alto, CA, USA, are all part of the San Francisco Bay Area. Second, business categories are self-reported and curated using a crowdsourcing approach. Each company contains multiple business categories. Third, the business description field contains rich textual, unstructured data about the organization. Given our interest in understanding venture positioning in global entrepreneurial ecosystems, our aim was to constrain our analysis to a select set of geographically-defined regions with particularly active levels of entrepreneurship. We based our selection on an assessment of well-represented regions in the Crunchbase dataset as well as most prominent startup regions identified in established industry reports (e.g., Compass Report and the 2017 Startup Genome Report). We intentionally did not limit ourselves to North American and European regions, but also included prominent entrepreneurial ecosystems in South America, the Middle East, Asia, and Australia for a truly global comparison. Our final region list contains a diverse, global set of 35 established and emerging entrepreneurial ecosystems, shown in Table I. The list contains 11 U.S. ecosystems and 24 non-U.S. We corroborated the validity and coverage of this list with venture capital investors and entrepreneurs. With a focus on these 35 global entrepreneurial ecosystems, our sample size reduced to 30 081 companies. We dropped companies with no business description and/or business categories. Our final sample ultimately consisted of 24 068 operating ventures in 35 regions. Table I provides descriptive statistics of our sample by entrepreneurial ecosystem.

B. Data Mining and Analysis

1) Text Mining of Venture Position Statements: The company description field in Crunchbase contains information on a venture’s activities as well as the particular value it provides to various stakeholders. More broadly, a company description can be considered a proxy for a venture’s position statement as outlined earlier. Company descriptions are provided in natural language form. In order to convert this data into useful information, text analytic methods must be applied. All text analytic methods convert natural language text blocks into a set of words. Following convention, we first remove all stop words. Since not all words in a position statement carry the same level of importance, we employ a weighted scheme to the set of extracted words. With regard to the distinguishing power of a term, we face a tradeoff relationship in term frequency. The more frequent a term appears in a statement, the more likely the term is characterizing some piece of information for that statement. However, if the term appears across all statements, it loses its distinguishing power. In order to balance term frequency within and across statements, we utilize term frequency-inverse document frequency (TF-IDF), a well-established weighting method, which normalizes term frequency by the rarity of the term across all documents. We use the gensim Python library for the implementation of the algorithm.
More formally we compute

\[ w_{t,d} = tf_{t,d} \times idf_{t,D} = \frac{freq_{t,d}}{\log_2 |D|} \]

(1)

where \( t \) is the focal term, \( d \) is the focal document, and \( D \) is the total set of documents. \( nt \) is the number of documents that contain term \( t \). \( tf_{t,d} \) stands for “term frequency of \( t \) in \( d \)” and \( idf_{t,D} \) stands for “inverse document frequency of \( t \) in \( D \).” According to this equation, the term frequency is computed as the raw count of the term \( t \) in the document \( d \) and the inverse document frequency is the logarithm of the inverse ratio of the number of documents that contains the term \( t \). Thus, this weight is a combination of local measure (i.e., term frequency) and global measures (i.e., inverse document frequency).

2) Computation of Venture Similarities: After converting the textual content into weighted vectors, we can now quantitatively compare position statements between any two ventures in an entrepreneurial ecosystem using the cosine similarity measure as follows:

\[ \cos(w_p, w_q) = \frac{w_p \cdot w_q}{\|w_p\| \cdot \|w_q\|} \]

(2)

where \( w_p \) and \( w_q \) are the normalized weighted vectors of startup \( p \) and \( q \), respectively. Fig. 2 provides an illustrative example of this computation. We perform this similarity computation for each pair of companies within each of the 35 entrepreneurial ecosystems. Since the number of pairs \([N(N-1)/2]\) grows quadratically to the number of ventures \( N \), we utilized high-performance computing infrastructure to accelerate the process. It should be noted that there are several ways of conceptualizing similarity between ventures, including structural, funding, leadership, or performance. Our aim in this paper was not to develop a comprehensive venture similarity vector, but instead focusing on the similarity in venture descriptions.

3) Construction of Venture Similarity Network: The venture (position statement) similarity network consists of nodes (representing ventures) and links (drawn between two nodes if there is a similarity between two ventures). This resulted in an almost fully connected network in all global ecosystems. An examination of pairwise similarities, however, revealed that similarities ranged from 0–100% and exhibited a highly-skewed distribution to the right. This suggests that more than 90% of venture pairs are in fact not significantly similar to each other. To reduce the density and generate a more manageable network for subsequent analysis and visualization, we used a conservative link similarity threshold of 15% to eliminate insignificant links between ventures. Our choice of using this similarity threshold was grounded in the prior work in text mining (e.g., [58] and [59]) and experimentation and expert interpretation with different threshold levels.

4) Modularity-Based Cluster Analysis: With each entrepreneurial ecosystem represented as a network, many different structural characteristics can be computed to build an understanding regarding the dynamics of the ecosystem. As we are interested in the macroscopic composition of venture positioning in each entrepreneurial ecosystem, detection of clusters and subcommunities is of importance [60], [61]. One prominent approach to identify clusters is to compute the modularity of the network [62]. Modularity measures the strength of division of
a network into modules (or groups, clusters, or communities) [63]. In other words, it enables to determine whether there are some structurally-induced communities in the ecosystem, rather than using traditional groupings such as industry category.

Given a partition of a network, modularity is computed as the difference between a fraction of links that connect nodes in the same group and the expected fraction if links were generated at random [63]. Modularity is, thus, maximized when a proposed partition gathers nodes that are densely connected to each other and separates out nodes that are not linked frequently. Formally, following [64], modularity for a given ecosystem is defined as follows:

$$M = \sum_{i=1}^{|C|} \left( \frac{l_i}{E} - \frac{e_i}{E} \right)^2$$

(3)

where \(|C|\) is the total number of identified clusters, \(l_i\) is the number of links within the \(i\)th cluster, \(e_i\) is the number of all links (local and bridging) connecting to ventures in the \(i\)th cluster, and \(E\) is the total number of ecosystem connections.

While several different modularity algorithms exist, we chose to use Louvain’s modularity-based clustering algorithm [62] to identify the position statement community structure of each entrepreneurial ecosystem due to its ability and performance to detect communities in large graphs as well as its implementation in our visualization software.

5) Topic Modeling: While the computation of the clusters provides an understanding of the overall community structure of an ecosystem, it provides little insight into what characterizes a cluster. To provide textual labels for each cluster, we again leverage the position statements of ventures within a given cluster, and utilize topic modeling techniques to provide human-interpretable labels.

The weight of keywords that best describe an ecosystem clusters are again computed using the TF-IDF method. One adjustment we make for cluster labeling is that we regard the company descriptions of the entire set of ventures as a document rather than an individual company description. More formally we compute

$$w_{t,c} = \text{tf}_{t,c} \times \text{idf}_{t,C} = \text{freq}_{t,c} \times \log_2 \frac{|C|}{n_t}$$

(4)

where \(t\) and \(c\) are the focal term and the focal cluster, \(|C|\) is again the total number of identified clusters and \(n_t\) is the number of clusters that contain term \(t\). We give a higher weight to the terms that appear frequently in a given cluster, but the weight is lowered if the term appears in many other clusters as well. After assigning the weight to each term for a given cluster, we extract the top five keywords based on the TF-IDF weight. The five keywords combined together describe a particular cluster in a given ecosystem.

6) Metrics: In addition to modularity, we compute a number of other well-established network statistics for each entrepreneurial ecosystem, also shown in Table I. The first measure—nodes—is a count of all the ventures in a given ecosystem. The next two metrics provide the number of nodes and links in the largest (main) component of the ecosystem, respectively. Components refer to the connected set of nodes in a network. The main component refers to the largest such connected component in the ecosystem network. All ratio columns in the table refer to the proportion of nodes and links to that of the preceding column.

We also computed a set of commonly used entrepreneurial ecosystem measures, including the average number of funding rounds, the average amount of funding (in $ million), the average amount of funding per round (in $ million), and the number of unique industries (a count of the unique number of categories associated with company profiles in each ecosystem). We also computed industry diversity using a modified Blau Index [65] as follows:

$$\text{Blau Index} = 1 - \text{HHI} = 1 - \sum_j p_j^2$$

(5)

where \(p_j\) is the proportion of category \(j\) in the total population of the ecosystem. We use the top 20 categories to compute sensible values for the Blau index. As a robustness check, we compute
the Blau index using the top 10 categories and all categories as well. The pairwise correlation among three measures are above 0.85. Categories are weighted as $1/n_i$ for company $i$ when it has $n_i$ categories.

Table I also presents the most frequently occurring industry in each ecosystem and proportion of that category in the ecosystem. When computing the top industry, we apply the same category weights.

The mean similarity column shows the mean of the pairwise similarities among the ventures in the ecosystem as a percentage.

C. Visualization

Visual representations are a fundamental component of human learning and understanding [66]. They enable us to not only communicate information or facts but also create, assess, and transfer insights, experiences, expectations, and perspectives. There is no algorithmic approach of choosing a single best visual representation for a given dataset; instead, there are many different representations available, each useful under different conditions and with its own advantages and limits [57], [67]. The choice is generally guided by the nature of the underlying data and the questions that are being asked. Given that startup position statement similarity and community are key issues of our inquiry, visual representations that can depict interconnectivity, positions, and clusters are particularly suitable [68].

One prominent approach is to use force-directed network layouts, which arrange nodes based on laws of attraction and repulsion from classical physics [69]. The use of a force-based layout is particularly appealing when the motivating issue is to identify central or prominent nodes, peripheral actors, and clusters. While there are many variations (e.g., Force-Atlas, Kamada Kawai, Yifan Hu, Fruchterman-Reingold [70]), we ultimately chose the OpenORD graph layout [71] implemented in Gephi [72]. The advantage of the OpenORD algorithm is that it allows specification of five distinct phases that enable cluster-differentiating layouts. As it is possible that some nodes can overlap with each other, we also applied the nooverlap algorithm to minimize occlusion.

In order to make visualizations readable, effective, and memorable, appropriate visual encodings must be selected. Visual
encoding refers to the selection of graphical properties of data marks for the network primitives. There are many different ways to visually encode nodes and links and the choice often depends on the underlying data type and tasks [73]. Nodes are depicted using a circle, colored by cluster membership, and proportionally sized by their prominence (as measured by closeness centrality). We use curved rather than straight links to reduce clutters caused by link crossings; the thickness of a link is proportional to the similarity of two ventures, and the color is based on the interpolation of the two nodes.

IV. RESULTS AND DISCUSSION

Table I shows a summary of metrics computed for all entrepreneurial ecosystems. The measures highlight that there are wide differences in the size, structure, diversity, and positioning of global entrepreneurial ecosystems. Ecosystems range in size (in terms of number of operating ventures) from the low hundreds (São Paulo, Brazil, with 114 ventures) to the thousands (Silicon Valley, USA, with 6297 ventures). The average ecosystem size is 687, but it is highly skewed by the existence of a few very large ecosystems, as indicated by the high variance. Not surprisingly, some of the largest entrepreneurial ecosystems are based in North America, including New York City, Boston, and Los Angeles. The largest non-U.S. entrepreneurial ecosystem is located in London (England) followed by Toronto (Canada), Tel Aviv (Israel), and Beijing (China).

The top tag-based industries across these entrepreneurial ecosystems include software (21) followed by e-commerce (6), biotechnology (5), and mobile (3). Interestingly, in U.S. entrepreneurial ecosystems where biotechnology is the top industry, it occupies a disproportionate presence in comparison to other industries and the overall industry diversity is low. For instance, in San Diego, biotechnology represents 25.06% of firms (compared to a global average of 10.86%) and industry diversity is significantly below average at 0.8326 (global average of 0.9120). This same pattern holds true for Boston and Raleigh.

Given the scale of our venture similarity networks, we followed scholarly convention to focus on the main component of each ecosystem for our core analyses and visualizations. The main component captures the largest group of ventures which are linked by direct or indirect ties (one venture is similar to another, which is similar to a third). In this component, all ventures are linked to each other in some way, and any one can be related to any other via a similarity pathway. As our results show, main component sizes range from very small (Munich) to very large (Silicon Valley). Indeed, the ratio of nodes in main components to total nodes ranges from 4.76% to 80.23%, with a global average of 30.64%. We observe very similar proportions when examining links in the overall ecosystem as well as the
main component. While we apply 15% as the cutoff threshold for similarities among ventures, a different cutoff threshold will likely change the size of main components in each ecosystem. The higher the threshold is, the more fragmented the ecosystem network will be, and the smaller the main component size will be.

The number of clusters, reflecting the number of strategic position groups in an entrepreneurial ecosystem, ranges from 2 to 30, and a global average of 12.30. The number of unique industries ranges from 158 (in Mumbai) to 723 (in Silicon Valley). Both of these measures are highly positively correlated with ecosystem size.

The average number of funding rounds ranges from 1.63 to 1.93, with a global mean of 1.78. The average funding amount (in $M) ranges from $8.04 (in Dublin) to $21.49 million (in Washington DC). The average funding amount per round ranges from $3.26 (in Munich) to $13.61 million (in Washington DC).

While the funding figures for DC are disproportionately high, and potentially an outlier, it is most likely explained by the resource-intensive defense and security industry that is present in this ecosystem. Indeed, when examining the visual representation of the DC ecosystem, the most central clusters include security and defense.

While visual representations of business ecosystems have been increasing in recent years [69], [74], [75], most studies focus on relatively well established interfirm relationship types between stakeholders (e.g., partnerships, investments, customer/supplier, people placements). What is much rarer is to derive ecosystem relationships between entrepreneurial ventures based on similarity of textual content. Our study fills this gap.

We visualized each of the entrepreneurial ecosystems using a network visualization approach. Visual representations not only help communicate complex network structures but also facilitate both comprehension and sense making. Given the size of some of the ecosystems, visual representations are particularly helpful.

Consider the visualization provided in Fig. 3. It presents a visualization of the structure of venture similarity in the Silicon Valley ecosystem. Nodes represent ventures; links represent similar position statements used by ventures. Positioning statement clusters are grouped by color. As we are using a cluster-emphasizing layout algorithm, our visualizations can be
interpreted as follows. First, similar nodes (ventures) typically cluster together. Centrally located nodes are core to the ecosystem, indicating usage of core and bridging strategic positioning language. Peripheral clusters are less central and tend to use more unique strategic positioning statements. Clusters that are closer to each other indicate a high number of overlap and interrelated statements; similarly, greater distance between clusters indicate a low number of overlap. Dense clusters contain highly similar strategic position statements; spread out clusters use highly differentiated language. Lastly, a node bridging two clusters can indicate that a startup potentially synthesizes disparate strategic positioning ideas.

With these interpretation guidelines in mind, we observe that Silicon Valley contains a rich set of unique strategic positioning clusters. Some of these clusters are centrally located, while there are several located peripherally. We also notice some very dense clusters, while others are spread apart. We observe a variety of such patterns across our entrepreneurial ecosystems.  

Intuitively, one may think that each cluster represents a specific industry. However, when examining Fig. 4, which depicts a zoomed-in version of the London ecosystem, we can see that there are many different tag-based industry categories within a cluster. In this example, ventures associated with analytics, mobile, design, payments, finance, and hospitality form the blue cluster. This result provides strong evidence that the clusters of strategic positions we uncover are not simply based on industry classification. In fact, it demonstrates that when we allow

---

9Given the number of visualizations, we only present a subset of illustrative examples here. A complete list of visualizations is provided on request.
these positions and clusters to emerge endogenously rather than through a particular industry classification, novel value configurations appear.

To gain a deeper understanding what the clusters actually mean, we applied topic modeling to each of the clusters within each entrepreneurial ecosystem. Consider the Boston ecosystem visualized in Fig. 5. First, we notice a centrally located cluster, described by words such as backup, storage, security, and virtualization. Boston is well known for its enterprise and security industries, and it appears that their strategic positioning statements are core to the overall ecosystem. We also notice a tightly integrated cluster of energy companies, related to wind, energy, cooling, clean, and vehicle, suggesting the positioning similarity between the energy and automotive industries. What is particularly interesting is the six labeled clusters at the bottom of the ecosystem visualization. Boston’s largest industry is biotechnology, but we see a nuanced differentiation in strategic positioning. One cluster appears to be related to the patient care and delivery facing side, while others are focused on specific subfields, including therapeutic treatments, cancer, DNA sequencing, and robotics. The proximity of these clusters confirm that they are related, but the varying topic labels reveal that there are some important differences.

With an understanding that ecosystems are characterized by the presence of venture similarity clusters, it is pertinent to explore how this measure relates to other metrics and what differences across ecosystems exist. Fig. 6(a)–(b) present how ecosystem size correlates with the number of clusters and the mean similarity in an ecosystem. Given the skewed distribution of ecosystem size, we utilized a logarithmic scale for the x-axis. Several interesting observations can be made. For instance, Fig. 6(a) compares the number of clusters against the ecosystem size at our default threshold level (15%). It clearly shows a strong positive correlation between the two measures. Since the x-axis is in log scale, the number of clusters identified is quickly saturating as the ecosystem grows in size.

Fig. 6(b), on the other hand, shows the comparison between the overall mean similarity and the ecosystem size. Here, we observe that the mean similarity of position statements for small size ecosystems has a much wider range. As ecosystems grow in size, the strategic positioning similarity of firms within that ecosystem converges to the global mean. This result is interesting as it does not necessarily have to be the case. On the one hand as an ecosystem grows in size, it becomes attractive to a more diverse set of firms. On the other hand, competition grows, and hence, strategic positioning become more similar.

Despite this convergence to the mean, we do however observe that there are still significant differences between positioning statements within an entrepreneurial ecosystem as demonstrated by the number of clusters. It is, thus, evident that ventures are continuously trying to balance legitimacy and uniqueness.

V. CONCLUDING REMARKS

This study defined and applied a rich, data-driven analysis and visualization approach for understanding the structure of venture similarity in global entrepreneurial ecosystems. Fusing data mining, text analytics, and network visualization, we examine the structure of strategic positioning of nearly 60,000 ventures in 35 ecosystems. Our visual analysis reveals that there are wide differences in entrepreneurial ecosystem size, structure, diversity, and positioning.

Specifically, we find that for entrepreneurial ventures in larger ecosystems, similarity in positioning statements converges to the global mean. We also find that ecosystems located in emerging economies tend to be smaller and have greater dispersion of venture similarity. That is to say, emerging ecosystems tend to have firms that position themselves either very similarly or very differentiated. The implication is that ventures within entrepreneurial ecosystems are constantly balancing legitimacy and differentiation and that balance becomes more salient as the ecosystem grows in size. Our ecosystem visualizations reveal that clusters are composed of firms from diverse industries and that some clusters are more proximate to each other than to others, suggesting potential differentiable intra- and inter-industry dynamics.

While our results are predominantly descriptive in nature—a natural outcome of visual exploratory analysis—our unique data-driven approach provides an important foundation for exciting future technology, innovation, and entrepreneurship research. First, we demonstrate that a data-driven approach provides the ability to study entrepreneurial ecosystems at scale. Second, we show that visualization is largely about hypothesis-generating, rather than hypothesis-testing. We anticipate interesting new empirical studies emerging from our paper. Associating the structural characteristics (e.g., cluster positioning and density) to the financial performance and health of an entrepreneurial ecosystem, for instance, is of great interest to both scholars and practitioners. Another interesting extension of our paper is to understand the evolution of strategic positioning in ecosystems and identifying potential mimetic processes across ecosystems. Given the generalizability of our data-driven approach, another promising avenue for future research would include an examination of entire actual business model descriptions, press releases, and patent announcements of ventures.

Our study also has important managerial implications. The visual text-analytic approach presented in this study provides industry analysts and policy makers a new methodology with which to conduct competitive market intelligence, which could be applied and extended to many other contexts. Our specific results will help entrepreneurs and investors understand the varied nature of entrepreneurial ecosystems, identify venture similarity clusters, and aid in devising effective strategic positions.

We acknowledge that our study is not without limitations. We use a socially-curated dataset, which may have some quality issues, such as missing or outdated data. We tried to overcome this by extracting the most up-to-date dataset and performing manual spot checks. Second, Crunchbase is a predominantly technology-industry oriented data source. While a wide variety of industries are present, it is quite possible that some industries are underrepresented or missing. Lastly, for manageability of analysis and visualization, we only investigated 35 global.

10 Appendix A shows the results of our sensitivity analysis at different threshold levels.
TABLE II

ECOSYSTEM CLUSTERS AT DIFFERENT THRESHOLD AND RESOLUTION LEVELS

<table>
<thead>
<tr>
<th>Region</th>
<th>Country</th>
<th>Threshold (%</th>
<th>Resolution</th>
<th>0.87</th>
<th>0.85</th>
<th>0.83</th>
<th>0.81</th>
<th>0.79</th>
<th>0.77</th>
<th>0.75</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>USA</td>
<td>5.00</td>
<td>5.66</td>
<td>6.00</td>
<td>6.00</td>
<td>6.00</td>
<td>6.00</td>
<td>6.00</td>
<td>6.00</td>
<td>6.00</td>
</tr>
<tr>
<td></td>
<td>USA</td>
<td>6.00</td>
<td>6.00</td>
<td>6.00</td>
<td>6.00</td>
<td>6.00</td>
<td>6.00</td>
<td>6.00</td>
<td>6.00</td>
<td>6.00</td>
</tr>
<tr>
<td></td>
<td>USA</td>
<td>12.50</td>
<td>12.23</td>
<td>11.98</td>
<td>11.74</td>
<td>11.49</td>
<td>11.35</td>
<td>11.22</td>
<td>11.28</td>
<td>11.35</td>
</tr>
<tr>
<td></td>
<td>USA</td>
<td>9.16</td>
<td>8.94</td>
<td>8.75</td>
<td>8.58</td>
<td>8.42</td>
<td>8.27</td>
<td>8.13</td>
<td>8.24</td>
<td>8.39</td>
</tr>
<tr>
<td></td>
<td>USA</td>
<td>8.08</td>
<td>7.89</td>
<td>7.72</td>
<td>7.57</td>
<td>7.43</td>
<td>7.30</td>
<td>7.18</td>
<td>7.29</td>
<td>7.45</td>
</tr>
</tbody>
</table>

TABLE III

SPEARMAN RANK CORRELATIONS FOR COMBINATIONS OF SIMILARITY THRESHOLD AND RESOLUTION

<table>
<thead>
<tr>
<th>Similarity Threshold</th>
<th>Resolution</th>
<th>N</th>
<th>ρ</th>
</tr>
</thead>
<tbody>
<tr>
<td>10%</td>
<td>-0.5</td>
<td>35</td>
<td>0.9574***</td>
</tr>
<tr>
<td>10%</td>
<td>-0.4</td>
<td>35</td>
<td>0.9258***</td>
</tr>
<tr>
<td>10%</td>
<td>-0.3</td>
<td>35</td>
<td>0.9112***</td>
</tr>
<tr>
<td>10%</td>
<td>-0.2</td>
<td>35</td>
<td>0.9123***</td>
</tr>
<tr>
<td>10%</td>
<td>-0.1</td>
<td>35</td>
<td>0.8686***</td>
</tr>
<tr>
<td>10%</td>
<td>0</td>
<td>35</td>
<td>0.8219***</td>
</tr>
<tr>
<td>10%</td>
<td>0.1</td>
<td>35</td>
<td>0.8011***</td>
</tr>
<tr>
<td>10%</td>
<td>0.2</td>
<td>35</td>
<td>0.8437***</td>
</tr>
<tr>
<td>10%</td>
<td>0.3</td>
<td>35</td>
<td>0.8272***</td>
</tr>
<tr>
<td>10%</td>
<td>0.4</td>
<td>35</td>
<td>0.8571***</td>
</tr>
<tr>
<td>10%</td>
<td>0.5</td>
<td>35</td>
<td>0.8429***</td>
</tr>
<tr>
<td>15%</td>
<td>-0.5</td>
<td>35</td>
<td>0.9496***</td>
</tr>
<tr>
<td>15%</td>
<td>-0.4</td>
<td>35</td>
<td>0.9574***</td>
</tr>
<tr>
<td>15%</td>
<td>-0.3</td>
<td>35</td>
<td>0.9633***</td>
</tr>
<tr>
<td>15%</td>
<td>-0.2</td>
<td>35</td>
<td>0.9465***</td>
</tr>
<tr>
<td>15%</td>
<td>-0.1</td>
<td>35</td>
<td>0.9536***</td>
</tr>
<tr>
<td>15%</td>
<td>0</td>
<td>35</td>
<td>0.9547***</td>
</tr>
<tr>
<td>15%</td>
<td>0.1</td>
<td>35</td>
<td>0.9541***</td>
</tr>
<tr>
<td>15%</td>
<td>0.2</td>
<td>35</td>
<td>0.9642***</td>
</tr>
<tr>
<td>15%</td>
<td>0.3</td>
<td>35</td>
<td>0.9643***</td>
</tr>
<tr>
<td>15%</td>
<td>0.4</td>
<td>35</td>
<td>0.9431***</td>
</tr>
<tr>
<td>15%</td>
<td>0.5</td>
<td>35</td>
<td>0.9355***</td>
</tr>
<tr>
<td>20%</td>
<td>-0.5</td>
<td>35</td>
<td>0.83***</td>
</tr>
<tr>
<td>20%</td>
<td>-0.4</td>
<td>35</td>
<td>0.8249***</td>
</tr>
<tr>
<td>20%</td>
<td>-0.3</td>
<td>35</td>
<td>0.7995***</td>
</tr>
<tr>
<td>20%</td>
<td>-0.2</td>
<td>35</td>
<td>0.7928***</td>
</tr>
<tr>
<td>20%</td>
<td>-0.1</td>
<td>35</td>
<td>0.7501***</td>
</tr>
<tr>
<td>20%</td>
<td>0</td>
<td>35</td>
<td>0.7473***</td>
</tr>
<tr>
<td>20%</td>
<td>0.1</td>
<td>35</td>
<td>0.744***</td>
</tr>
<tr>
<td>20%</td>
<td>0.2</td>
<td>35</td>
<td>0.741***</td>
</tr>
<tr>
<td>20%</td>
<td>0.3</td>
<td>35</td>
<td>0.7289***</td>
</tr>
<tr>
<td>20%</td>
<td>0.4</td>
<td>35</td>
<td>0.7293***</td>
</tr>
</tbody>
</table>

Note: ρ is the Spearman rank correlation. *** denotes p-value less than 0.1%

APPENDIX

A. Similarity of Company Descriptions

To assess the quality of business descriptions, we compared the business description provided by Crunchbase with actual corporate website descriptions for 50 randomly selected firms. Specifically, we used a TF-IDF approach to compare the similarity between two descriptions. If the TF-IDF score is greater or equal to 0.15, the two descriptions are deemed similar. This follows the approach we have used throughout the paper and what is recommended by the prior work. Given space limitations, the detailed results of our analysis are provided in an online appendix.11 The descriptions from two distinct sources are on average 53.89% similar and three companies have 100% similar descriptions on both the corporate website and Crunchbase. The lowest similarity level is still 19.54%. These results

11http://entsci.gatech.edu/venturesimilarity/e_supplement.pdf
suggest significant content overlap with the Crunchbase entry, giving us confidence about the validity of the data.

B. Sensitivity Analysis

We conducted several sensitivity analyses to ensure the robustness of our results. Specifically, for our network modeling part, we conducted analyses at three different similarity threshold levels (10%, 15%, and 20%) as well as a ±0.5 variation (using 0.1 increments) on the resolution parameter (see [76]) of the Louvain modularity algorithm. The three threshold levels represent a broad spectrum of both more relaxed and stringent similarity criteria. A 10% threshold level suggests that only 10% of the descriptions need to match in order for a link to be drawn between two firms; a 20% threshold on the other hand suggests that at least 20% of the descriptions need to match.
For each configuration of threshold and resolution, we report the mean after 100 replications (see Table II). The sensitivity analysis was implemented using Python\textsuperscript{12} and NetworkX.\textsuperscript{13}

Our results show that increasing the threshold level reduces the number of clusters, in general, which in turn increases the number of clusters as the network gets more disconnected (i.e., greater threshold $\rightarrow$ less similarity $\rightarrow$ reduction in edges $\rightarrow$ increase in clusters). Fig. 7 shows the results of our sensitivity analysis with respect to the association between ecosystem size and number of clusters differentiated by the threshold level [corresponding to Fig. 6(a)]. We observe that for a lower threshold level the number of clusters increases for the majority of ecosystems (25/35). As the threshold level increases, we see a lower number of clusters in the main component of each ecosystem. However, this finding is less pronounced and in fact in some instances opposite for larger ecosystems as higher threshold levels do not reduce the main component in significant ways as for smaller ecosystems. We, thus, observe that for ecosystems with large main components (e.g., London, New York, or Boston) that the number of clusters initially increases and then decreases with higher threshold levels. While increasing the threshold level will by definition improve the similarity between companies, we lose important information that may exist in the network. We, thus, adopt the 15\% threshold level for our main analysis.

To examine the positive relationship between the ecosystem size and the number of clusters, we created small multiples for each combination of resolution and threshold level as shown in Fig. 8. We also computed the Spearman rank correlation to see how stable the correlation is for different levels of resolution. Overall, the Spearman correlation for the whole aggregate data is $0.5995$ for $N = 1,155$. Next, we computed the correlations for each combination of similarity threshold and resolution (see Table III). Together, the correlations from the visuals and analysis confirm the positive relationship between ecosystem size and the number of clusters from the main component.

\begin{thebibliography}{99}
\bibitem{7} V. W. Hwang and G. Horowitz, The rainbowforest: The secret to building the Next Silicon Valley. Regenwald Los Altos Hills, CA, USA: Createspace Independent Pub., 2012.
\bibitem{36} A. Saxenian, Regional Advantage. Cambridge, MA, USA: Harvard Univ. Press, 1994.
\bibitem{38} D. Isenberg, “The entrepreneurship ecosystem strategy as a new paradigm for economic policy: Principles for cultivating entrepreneurship,” Presentation at the Institute of International and European Affairs, Dublin Ireland, May 12, 2011.
\end{thebibliography}