MIS Quarterly

BIG DATA & ANALYTICS IN NETWORKED BUSINESS

TOWARD A BETTER MEASURE OF BUSINESS PROXIMITY: TOPIC MODELING FOR INDUSTRY INTELLIGENCE[1]

Zhan (Michael) Shi
Department of Information Systems, W. P. Carey School of Business, Arizona State University,
Tempe, AZ 85287-4606 U.S.A. {zmshi@asu.edu}

Gene Moo Lee
Department of Information Systems and Operations Management, College of Business, The University of Texas at Arlington,
Arlington, TX 76019 U.S.A. {gene.lee@uta.edu}

Andrew B. Whinston
Department of Information, Risk, and Operations Management, McCombs School of Business, The University of Texas at Austin,
Austin, TX 78712 U.S.A. {abw@uts.cc.utexas.edu}

In this article, we propose a new data-analytic approach to measure firms' dyadic business proximity. Specifically, our method analyzes the unstructured texts that describe firms' businesses using the statistical learning technique of topic modeling, and constructs a novel business proximity measure based on the output. When compared with existent methods, our approach is scalable for large datasets and provides finer granularity on quantifying firms' positions in the spaces of product, market, and technology. We then validate our business proximity measure in the context of industry intelligence and show the measure's effectiveness in an empirical application of analyzing mergers and acquisitions in the U.S. high technology industry. Based on the research, we also build a cloud-based information system to facilitate competitive intelligence on the high technology industry.

Keywords: Big data analytics, business proximity, topic modeling, industry intelligence, information system

Introduction

Business proximity measures firms' relatedness in the spaces of product, market, and technology, which is an important concept in industry intelligence and also a central building block in many studies of firm strategy and industrial organization. Not surprisingly, prior studies in different management disciplines have used or developed a handful of measures of business proximity. One common practice has been to classify firms into industries and to operationalize business proximity as a binary variable that indicates common industry membership. Under this definition, two firms' businesses are either identical or completely different. A refined extension of the binary definition has been to better utilize the hierarchical information provided by some industry classification system, such as Standard Industrial Classification (SIC) or North American Industrial Classification System (NAICS). For example, in Wang and Zajac (2007), the similarity of two firms' businesses was determined by the number of common consecutive digits in their industry classification codes under NAICS. Since they used the first four digits in NAICS, the similarity quantity was one of five possible values: 0.00, 0.25,

[1]Bart Baesens, Ravi Bapna, James R. Marsden, Jan Vanthienen, and J. Leon Zhao served as the senior editors for this paper.

The appendix for this paper are located in the "Online Supplements" section of the *MIS Quarterly*'s website (http://www.misq.org).

0.50, 0.75, or 1.00. However, this measure is still discrete, and the level of granularity it can achieve is constrained by the industry classification system on which it depends. There are several other measures that were aimed at some specific aspect of firms' businesses, and they typically had much stronger data requirements. Stuart (1998), Mowery et al. (1998), and others constructed a "technological overlap" measure using data of firms' patent holdings. The closeness of a pair of firms was assumed to be proportional to the number of common antecedent patents cited. While this is an elegant, continuous measure in the technology space, it requires complete data on firms' patent portfolios and does not explicitly cover the product and market spaces. Mitsuhashi and Greve (2009) applied the Jaccard distance on firms' customer geographic regions in measuring "market complementarity." Likewise, this measure focuses only on the (geographic) market space and requires all relevant firms' customer geography data to be available.

While these measures have served the researchers' purposes well, we see an opportunity for a new and more general methodology in light of the increasing availability of public, unstructured data and recent advances in big data analytics. In this paper, we propose a method that requires little manual preprocessing yet provides finer granularity on quantifying firms' positions in the spaces of product, market, and technology. Utilizing a statistical learning technique called topic modeling (Blei 2012), we analyze the publicly available, unstructured texts that describe firms' businesses. Our automatic approach, the core of which is a Latent Dirichlet allocation (LDA) algorithm, represents each firm's textual description as a probabilistic distribution over a set of underlying topics, which we interpret as aspects of its business. The data-analytic framework greatly reduces the complexity of representing the business environment, and produces structured information that enables further examination and derivation. Our new business proximity measure is then naturally constructed by quantifying the "distance" between a pair of firms' topic distributions.

An important advantage of our method for measuring business proximity is that it imposes a much weaker requirement on structured data than the existent measures. This makes our approach particularly appealing when the firms under study are small and privately held, for which detailed information on industry classification, patent holding, and product/customer is either very sparse or not available at all. Motivated by this advantage, we choose the U.S. high technology (high-tech) industry as the empirical context to demonstrate our approach. We collect data from CrunchBase, an open and comprehensive source for high-tech startup activity. For the majority of companies in our dataset, the standardized industry classification code is unavailable, and due to various strategic reasons, most do not disclose their customer information and key intellectual property, so the conventional methods for measuring business proximity cannot be operationalized. Using this dataset as an example, we detail the procedure of our data-analytic approach, and compute business proximity for each pair of the companies. We then show the validity and effectiveness of the new measure in the context of industry intelligence by (1) examining the relationships between business proximity and simple category classification, between business proximity and job mobility, and between business proximity and investment, respectively, and (2) using the measure in a novel empirical application of modeling matching of companies in mergers and acquisitions (M&As). Our comprehensive, continuous measure is an enabler in the analysis to show the nuanced relationship between M&A transactions and the firms' business similarity and complementarity. Methodologically, to recognize the increasingly networked business environment as well as to accommodate the relational nature of the matching data, we employ an innovative statistical framework called exponential random graph models (ERGMs) in the M&A analysis.

This research joins the rapidly growing stream of information systems literature that leverages newly developed data science techniques in examining big data for business analytics (e.g., Adomavicius and Tuzhilin 2005; Chen et al. 2012; Chiang et al. 2012; Ghose et al. 2012; Shi et al. 2014; Shmueli and Koppius 2011; Xu et al. 2014). Our research shows how big data analytics can potentially transform competitive intelligence, particularly for the high-tech industry, where recent years have seen an "entrepreneurial boom" characterized by the explosion of digital startups. Such explosion has made it ever more difficult to purely rely on individuals' industry knowledge to depict the rapidly changing landscape of the startup world. Our empirical analysis demonstrates the potential of extracting economically meaningful information from publicly available, unstructured data through large-scale computation as well as the value of the proposed business proximity measure as an important metric in the analytics of M&A matching and as a search tool for navigating the networked startup world. To further illuminate the practical implication of our data-analytic framework, we build an information system that allows managers and analysts to use business proximity to explore the competitive landscape of the U.S. high-tech industry. The back end of our system handles data collection, storage, and large-scale computation using a big data computation platform (Condor), NoSQL database technology (MongoDB), and various programming languages (Python, Scala). The front end of the system is hosted on Google's Cloud Platform and provides users an easy-to-use web interface. It is available to access at http://diamond.mccombs.utexas.edu/bizprox.

We organize the remainder of this paper as follows. To provide a context for describing the data-analytic method, we first introduce our dataset. We then elaborate the procedure for constructing our business proximity. Next, we demonstrate the validity and effectiveness of our measure. We subsequently describe the information system implementation. Finally, we discuss and conclude our paper.

Data

The dataset for demonstrating our methodology was collected from CrunchBase.[2] CrunchBase is an open and free database of high-tech companies, people, and investors. Regarded as the Wikipedia of the high-tech industry, it provides a comprehensive view of the "startup world." CrunchBase keeps track of the industry by automatically retrieving and extracting information from professionally edited news articles on technology-focused websites (e.g., TechCrunch and Business Insider). In addition, ordinary users can contribute to CrunchBase in a crowdsourcing manner. For quality assurance, each update is reviewed by moderators. Existing data points are also constantly reviewed by the editors. Compared with other high-tech-focused data vendors, CrunchBase has the advantage of more complete coverage on early-stage startups, especially those not (yet) funded by venture capitalists.

Data collection was carried out between April 2013 and April 2015. The companies and their information were collected at the beginning of the period. We limit our dataset to the U.S.-based companies and exclude those for which some basic information (e.g., founding date, business description) is missing. We further exclude companies that had already been acquired as of April 2013. The resultant dataset contains 24,382 companies, the vast majority of which are privately held, early-stage startups that are unclassified under SIC or NAICS. As of April 2013, 345 of the companies (1.41%) in the dataset were publicly traded,[3] and the median age of the whole sample was 5.66 years old. For each company, we also observe its headquarters location, industry sector (CrunchBase-defined category), (co)founders, board members, key employees, angel and venture investors that participated in each of its funding rounds, acquisitions, investments, and a business description. Confirming the common knowledge about the high-tech industry, we observe considerable geographic clustering. Figure 1(a) visualizes the spatial distribution of the companies using the headquarters-location data aggregated at the city level. The circles are centered at the cities and their radius is proportional to the number of companies. The major high-tech hub cities include New York City (8.08% of the companies), San Francisco (7.92%), Los Angeles (2.17%), Chicago (2.10%), Seattle (1.93%), Austin (1.84%), and Palo Alto (1.81%). At the state level, as shown in Figure 2(a), California leads with 34.72% of the companies, followed by New York (11.99%), Massachusetts (5.89%), and Texas (5.20%). We also observe a highly uneven distribution of companies across the 19 industry sectors (CrunchBase-defined categories). The leading sectors are "software" (19.23%) and "web" (17.13%), and the trailing sectors are "semiconductor" (1.00%) and "legal" (0.73%), as shown in Figure 2(b). In the dataset, the people's profiles also contain their past professional experiences. The unstructured, textual descriptions are mostly of short to moderate length, comprising one or more paragraphs on the key facts about the companies' products, markets, and technologies.

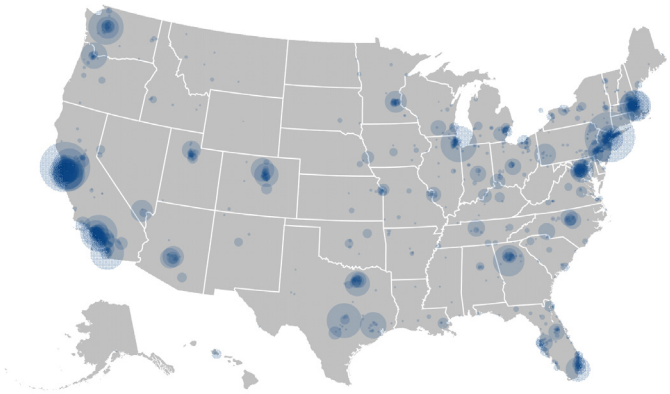
For the validation of the proposed method, we use three types of inter-firm interactions: M&A (one firm acquires another), investment (one firm invests in another), and job mobility (an individual changes jobs from one firm to another). We constantly monitored these activities to April 2015. Our dataset includes a total of 1,689 M&A transactions since 2008. Figure 1(b) geo-maps each of the M&A transactions using the headquarters locations of the involved companies. A little less than two-thirds (62.59%) of the deals are cross state. A numerically similar portion of transactions (63.56%) is cross sector. The distribution of the number of transactions per company is highly skewed—the top 10 and top 20 buyers made 14.32% and 21.23% of all the deals, respectively. Among these M&A transactions, 394 (23.32%) occurred between April 2013 and April 2015. For investments, a total of 531 transactions are recorded and the post-April-2013 number is 129 (24.29%). Finally, the job mobility data are computed based on position changes among the 24,334 people in the dataset. There are 19,697 company pairs connected by the job transitions in total and 9,792 pairs (49.71%) by post-April-2013 activities.

Measuring Business Proximity: Data-Analytic Framework

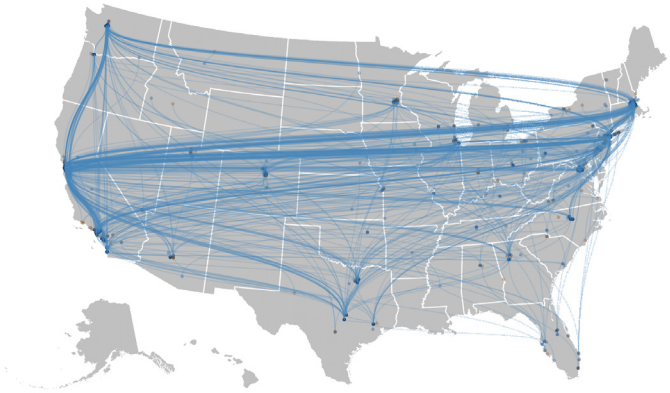
Business proximity measures firms' closeness in the spaces of product, market, and technology. Our objective is to develop a data-driven, analytics-based business proximity measure to improve on scalability, classification granularity, and comprehensiveness. The input of our method—an unstructured, textual business description for each firm—requires no manual classification, and is also much more likely to be

[2] http://www.crunchbase.com.

[3] Hence, financial statement information, such as SEC filings, is only available for a very small fraction of the companies in the dataset.

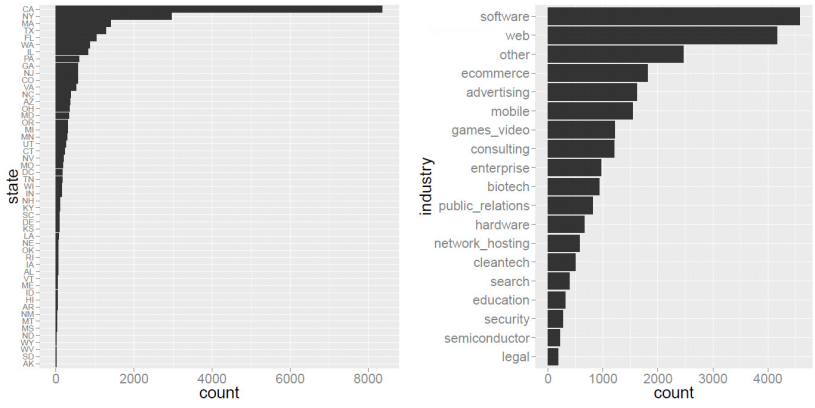


(a) Companies



(b) M&A Transactions

Figure 1. Geo-Mapping Company Locations and M&A Transactions



(a) State (b) Industry Sector

Figure 2. Distribution of Companies over State and Industry Sector

available than structured information such as NAICS/SIC code or patent portfolio, especially for high-tech startups.

Our approach builds upon a text mining technique called topic modeling, a statistical method that discovers abstract "topics" from a large collection of documents. At present, the most common topic modeling algorithm is Latent Dirichlet allocation (Blei et al. 2003). LDA does not require manually labeling each document, so it is an unsupervised learning algorithm. The underlying model of LDA is generative; the assumption is that each word in each document is probabilistically drawn from the vocabulary of a topic discussed in that document. Given a large collection of documents, the vocabularies of topics and the topics of the documents are jointly estimated.

More formally, we let the number of input descriptions (i.e., the total number of companies) be D, where each description $d \in \{1, 2, \ldots, D\}$ is a collection of words $\{w_n^d \mid n = 1, 2, \ldots, N^d\}$. Let the total number of latent topics (business aspects) expressed by the descriptions be K. Each topic $k \in \{1, 2, \ldots, K\}$ is a probabilistic distribution over the whole vocabulary (i.e., the set of unique words in the description corpus). This distribution is denoted φ^k, where φ_w^k is the probability of word w in topic k. The topic proportions for description d are θ^d, where θ_k^d is the topic proportion for topic k in description d. Assume z_n^d is the topic assignment of the n^{th} word in description d. Then, given θ^d and φ^k, the probability of observing description d is

$$\prod_{n=1}^{N^d}\left(\sum_{k=1}^{K}\mathrm{P}\left(w_n^d \mid z_n^d = k, \phi^k\right)\left(P\left(z_n^d = k \mid \theta^d\right)\right)\right) = \prod_{n=1}^{N^d}\left(\sum_{k=1}^{K}\phi_{w_n^d}^k \theta_k^d\right) \quad (1)$$

where the term inside the product operator is the probability of the n^{th} word in description d being w_n^d. LDA takes the Bayesian approach and is a complete generative model. It further assumes Dirichlet priors for both θ and φ, with hyperparameters α and β, respectively. Thus, the generative process of LDA can be represented by the following joint distribution:

$$\mathrm{P}(w, z, \theta, \phi \mid \alpha, \beta) = \prod_{k=1}^{K}\mathrm{P}\left(\phi^k \mid \beta\right)\prod_{d=1}^{D}\mathrm{P}\left(\theta^d \mid \alpha\right)\left(\prod_{n=1}^{N^d}\mathrm{P}\left(w_n^d \mid z_n^d, \phi^k\right)\mathrm{P}\left(z_n^d \mid \theta^d\right)\right) \quad (2)$$

Having observed the descriptions, hence w, we compute the posterior distribution

$$\mathrm{P}(z, \theta, \phi \mid \alpha, \beta) = \frac{\mathrm{P}(w, z, \theta, \phi \mid \alpha, \beta)}{\mathrm{P}(w \mid \alpha, \beta)} \quad (3)$$

using Monte Carlo methods in Bayesian statistics. Finally, the estimates of θ and φ are obtained by examining the posterior distribution.

In summary, LDA is utilized in our data-analytic framework to analyze the textual descriptions of the firms. Each description is a document, and all the descriptions together are the input of LDA. The algorithm produces K topics (K is a parameter specified by the researcher), each of which is represented by a probabilistic distribution over the set of words. In addition, LDA computes the topic distribution for each company description. For each company, a probability value, or weight, is assigned to each discovered topic and the values sum up to 1. Essentially, through topic modeling, company i's description is represented by a topic distribution $T_i = \{T_{i,1}, T_{i,2}, \ldots, T_{i,K}\}$ where $T_{i,k}$ is the weight on the k^{th} topic and $\sum_{k=1}^{K} T_{i,k} = 1$.

We interpret the discovered topics as the different components of the companies' businesses. If a particular $T_{i,k}$ has the value of 0, then component k is irrelevant to company i's business. Finally, we define the *business proximity* $p_b(i,j)$ between two companies i and j as the cosine similarity[4] of the two corresponding topic distributions T_i and T_j, which can be written as follows:

$$p_b(i, j) = \frac{T_i \bullet T_j}{||T_i||||T_j||} = \frac{\sum_{k=1}^{K} T_{i,k} T_{j.k}}{\sqrt{\sum_{k=1}^{K}\left(T_{i,k}\right)^2}\sqrt{\sum_{k=1}^{K}\left(T_{j,k}\right)^2}} \quad (4)$$

The resulting proximity values range between 0 and 1, where a bigger value indicates closer proximity between the pair of companies. The measure equals 0 if and only if the two firms have no common business component; the measure equals 1 if and only if the two firms share exactly the same business components as well as the same weights.

We carry out the proposed method on the CrunchBase dataset. We run the LDA model and compute the corresponding business proximity for a set of different K values: 50, 100, 200, and 500. The main results on coefficient signs and their statistical significance reported in the empirical validation and application section are robust to the different choices. Due to

[4]Cosine similarity is one measure of similarity between two distributions. We can apply other similiarity measures such as normalized Euclidean distance. We can also view each topid distribution as a set where the elements are the topics with strictly positive probability, and then use set comparison metrics such as Jaccard index and Dice's coefficient. Our main results are robust to these alternative measures.

Table 1. LDA Results of CrunchBase Data (Partial)*

Topic	Dimension	Top 5 Words
1	Product	video, music, digital, entertainment, artists
2	Product	news, site, blog, articles, publishing
3	Product	job, jobs, search, employers, career
4	Product	people, community, members share, friends
30	Technology/Product	phone, email, text, voice, messaging
31	Technology/Product	wireless, networks, communications, internet, providers
32	Technology/Product	cloud, storage, hosting, server, servers
33	Technology/Product	app, apps, iphone, android, applications
38	Market	sales, customer, lead, email, leads
39	Market	solution, cost, costs, applications, enterprise

*Only the top five words are presented for brevity.

the page limit, we report in the main text for $K = 50$. To illustrate that the topic model results comprehensively capture multiple dimensions of a firm's business, in Table 1 we list 10 topics that LDA produces from our dataset. Note that each topic is a distribution over all words in the vocabulary and that we only show the top five words in terms of their probability for brevity. The full 50-topic list is shown in Table A4 in the Appendix. We have checked all 50 topics to find that each topic consists of frequent words that are tightly related to each other. We also observe that the topics capture the current trends in the high-tech industry. Using the LDA results, we compute business proximity for all company pairs in the dataset. Owing to the huge number of pairs (close to 300 million), we parallelize the computation algorithm for speedy processing.

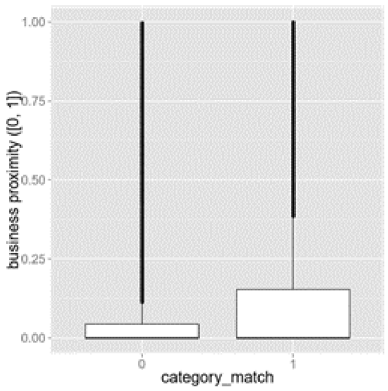
Our new data-driven approach for measuring business proximity has overcome many of the limitations faced by the existing methods. First, the approach is scalable because the construction of the business aspects and business proximity is automated, which is a sharp contrast to the domain-expert-based industry classification in which manual annotation is required as the first step. Second, our approach is generally applicable to a wide range of firms (either public or private) as long as textual business descriptions exist for the firms. In contrast, industry classification is only sparsely available for small companies and financial filings data are only available for public companies. Note that only 1.41% of the high-tech companies in our dataset are public, as discussed in the previous section. Third, our approach provides finer granularity than the existing discrete similarity measures as the algorithm provides continuous similarity measures. Fourth, the proposed method provides flexibility to cope with dynamic industry changes. As the underlying business descriptions in the industry change, the algorithm can automatically detect the emerging topics in the industry and incorporate them into the business proximity.

Empirical Validation and Application

Validation

To validate the constructed business proximity measure, we first examine the relationship between it and a simple category-based classification. Because the NAICS-based proximity cannot be operationalized due to the data limitation (in fact, the CrunchBase companies are already in a narrowly focused industry), we leverage the simple industry sector information, that is, the categories defined by CrunchBase (see Figure 2). We construct a binary indicator for same-category membership, category_match, and let it serve as a benchmark business proximity measure. We then compare the distributions of the proposed analytics-based measure in two groups of company pairs: (1) company pairs in the same category (category_match = 1), and (2) those belonging to different categories (category_match = 0).

Figure 3 compares the business proximity values between the two groups. The upper and lower hinges of the boxes indicate the first and third quartiles (the 25^{th} and 75^{th} percentiles). The results show that the same-category company group (mean: 0.12) has a mean business proximity value twice as large as the other (mean: 0.06). The Pearson's correlation coefficient between business proximity and category match is 0.11, with the t-statistic being 61.94 and p-value being smaller than $2.2e^{-16}$. The large t-statistic and low p-value indicate a very high correlation between the proposed business proximity and the simple category classification.



Note: The upper and lower hinges of the boxes indicate the 25th and 75th percentiles.

Figure 3. Distributions of Business Proximity: Same- and Cross-Category Company Pairs

For further validation, we test the predictive power of the proposed business proximity on three types of inter-firm interactions: M&A, investment, and job mobility.[5] Operationally, we compare the realized business proximity among four groups (M&A, invest, job mobility, and random) of company pairs to test if the business proximity has a leading effect on the corresponding inter-firm interactions. One caveat is that high business proximity values could be the result of firm transactions. For instance, after an M&A transaction takes place, it is very likely that the acquiring company's business description will incorporate various aspects of the acquired company. To avoid this reversal effect, we only consider the inter-firm transactions after April 2013, which is the time when all of the company descriptions were collected. Our inter-firm interaction dataset contains 394 company pairs associated to M&A transactions, 129 with inter-firm investments, and 9,792 with job mobility.[6] Finally, to construct the baseline, we randomly select company pairs from the whole dataset.

Figure 4 compares the distribution of business proximity value among the company pairs defined by M&A, investments, job mobility, and random selection. We find that the proposed business proximity has higher values between company pairs connected by the three types of inter-firm interactions than random pairs, thus indicating a positive association between each of the transactions and the proximity. On average, the first three groups have more than three times higher proximity than the randomly paired group: M&A (0.293), investments (0.224), job mobility (0.218), and random (0.068). Given the fact that M&A is a rare, significant inter-firm transaction, it is intuitive to find that M&A-paired firms have higher similarities than other two interaction types (investments and job mobility).

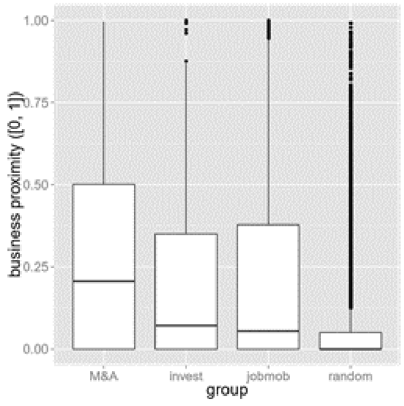
Empirical Applications on M&As

In this subsection, we demonstrate the business proximity measure's value for empirical modeling. Specifically, we apply it in analyzing high-tech M&As. Recognizing the increasingly networked business environment,[7] we construct a network structure by incorporating firm proximity in different dimensions, and then use a statistical network model to analyze their interactions. Our objective is to examine the relationship between the likelihood of a pair of firms' matching in an M&A transaction and their individual and pairwise characteristics, among which the newly developed business proximity is of our primary interest. We first summarize the theoretical basis for the importance of business proximity as well as proximity in three other dimensions in modeling M&As. Next, we introduce the statistical network analysis method and explain our empirical specifications. Finally, we present estimation results.

[5]The rationale of choosing these interactions is the following: M&A is an important inter-firm transaction that in theory creates business synergy (e.g., Rhodes-Kropf and Robinson 2008); inter-firm investments are associated with technological or market overlaps (e.g., Mowery et al. 1998), and may lead to future M&A transactions (Mikkelson and Ruback 1985); the labor economics literature found evidence that a significant portion of the job moves involve companies that are in the same industry (e.g., Fallick et al. 2006; Moscarini and Thompson).

[6]For job mobility, if a person made a job transition from company *A* to company *B*, then we consider *A* and *B* are associated.

[7]See "Revolution in Progress: The Networked Economy," *MIT Technology Review Custom*, August 27, 2014.



Note: The upper and lower hinges of the boxes indicate the 25th and 75th percentiles.

Figure 4. Distributions of Business Proximity: M&A, Investment, Job Mobility, and Random Samples

Proximity and M&A

The high-tech industry is characterized by active and rapid innovation, significant geographic clustering (at a handful of high-tech hubs), rapid job mobility, high concentration of ownership at the company level, and strong influence of angel and venture investors. We posit that business proximity, geographic vicinity, social linkage, and common ownership are associated with the likelihood of two firms' matching in an M&A transaction.

Business Proximity: Business proximity measures firms' relatedness in the spaces of product, market, and technology. It has been widely recognized in the finance and management literature that the potential synergy in products, markets, and technologies is a key driver for M&As (e.g., Rhodes-Kropf and Robinson 2008) and is especially important in high-tech acquisitions (e.g., Ahuja and Katila 2001). The central idea of business synergy is that economic surplus can be created from novel recombination of the acquirer's and target's resources and capabilities. One of the determinants for the matching of acquirer and target should be the recombination potential, which is in turn influenced by the relatedness of two firms' products, markets, and technology. Therefore, we expect the business proximity is associated with the M&A matching likelihood.

Geographic Proximity: Geographic or spatial proximity refers to the closeness of physical locations and it has been shown to have a moderating effect on a diversity of financial transactions. In the M&A domain, Erel et al. (2012) analyzed cross-border mergers to show that, among other factors, geographic proximity increases the likelihood of mergers between two countries. At the firm level, Chakrabarti and Mitchell (2013) found that chemical manufacturers prefer spatially proximate acquisition targets. The main reasoning behind these findings is that information propagation is subject to spatial distance; geographic proximity brings a higher level of knowledge exchange and hence a lower level of information asymmetry. For the same reason, we predict that geographic proximity is positively associated with the M&A likelihood.

We operationalize geographic proximity by measuring the great-circle distance[8] between two companies' headquarters. First, we translate the street address of each company's headquarters into its latitude (φ) and longitude (λ) coordinates by using Google Maps API.[9] For companies whose full street address is missing, we use the city center as an approximate. Next, we use the latitude and longitude coordinates to calculate the great-circle distance. Specifically, let (φ_i, λ_i) and (φ_j, λ_j) be the coordinates for companies i and j, and $\Delta\lambda$ be the absolute difference in their longitudes. Then the *geographic proximity* $p_g(i, j)$ between companies i and j is defined as

$$p_g(i,j) = -R \arccos\left(\sin\phi_i \sin\phi_j + \cos\phi_i \cos\phi_j \cos\Delta\lambda\right) \quad (5)$$

where the constant R is the sphere radius of the earth. The negative sign is to convert distance to proximity.

Social Proximity: Social proximity of two firms is defined according to the social linkage between the individuals associated with the two firms. Personal linkage is an impor-

[8]http://en.wikipedia.org/wiki/Great-circle-distance.

[9]https://developers.google.com/maps/.

tant factor in coordinating transactions and promoting private information exchange between business entities through mutual trust and kinship (e.g., Cohen et al. 2008; Hochberg et al. 2007; Stuart and Yim 2010). We believe two factors about the high-tech industry greatly contribute to the importance of personal linkage's role in transmitting vital information across companies. First, the U.S. high-tech industry, especially its startup sphere, is characterized by high job mobility, which creates the paths and opportunities for private information flow (Fallick et al. 2006). Second, early-stage digital startups are mostly very small in size; thus, information about them is often scarce outside the teams' social circles. Moreover, many startups intentionally stay in a "stealth mode" before their products and technologies mature. Hence, we argue that companies with closer social proximity are likely to be aware of each other's products and intellectual property, which would lead to a higher M&A probability.

We operationalize social proximity by using the "people" part of our dataset. For each company, we observe the individuals who are or have previously been affiliated with it either as a (co)founder, or as a board member, or as an employee. Let S_i denote this set of individuals for company i. Then we define the *social proximity* $p_s(i, j)$ between two companies i and j as

$$p_s(i, j) = |S_i \cap S_j| \tag{6}$$

that is, the number of people who are identified having experiences in both companies.

Investor Proximity: Investment proximity is defined according to the common angel and venture investors who have founded the firms. In the high-tech industry, startups depend on external investments to support product development before they establish a stable cash flow. As compared with other types of investors, angel and venture investors often play a more active role in management and can be highly influential on strategic decisions (e.g., Amit et al. 1990; Gompers 1995), such as pursuing M&A opportunities. Hence, common early investors of two high-tech companies can form a critical information bridge or even an initiator and enabler of collaboration between them, which we predict leads a higher likelihood of M&A.

Our operationalization of investor proximity is methodologically similar to that of social proximity. Given two companies i and j, their investor proximity $p_I(i, j)$ is defined as

$$p_I(i, j) = |I_i \cap I_j| \tag{7}$$

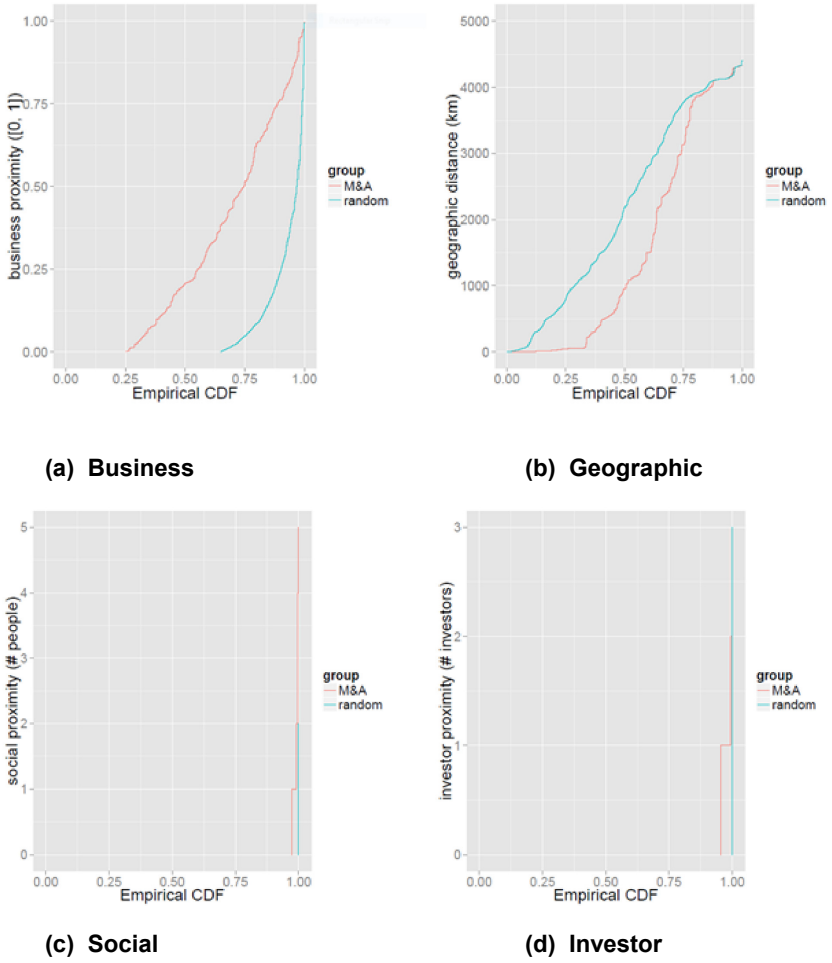
where I_i and I_j are the sets of investors who have funded companies i and j in any of the funding rounds respectively.

Correlation Analysis: We explore the realizations of the business, geographic, social, and investor proximities in our CrunchBase dataset and analyze their correlations with the matching of M&A. Note that we compute all proximity measures using company data collected in April 2013 and only use the M&A transactions that occurred between April 2013 and April 2015 to avoid any possible reversal effect.

For each of the four proximity measures, we compare its different distributions in two groups of company pairs: (1) group of M&A-matched company pairs and (2) that of randomly selected pairs. Figure 5 shows the empirical cumulative distribution functions (CDF) of the four proximity measures. For the geographic dimension, we plot the distance rather than proximity for intuitiveness. Also note that the business and geographic proximity values are continuous, whereas the other two are discrete. In each subfigure, the red line represents the distribution for the group of company pairs defined by M&A transactions and the green line shows that of random pairs. For each proximity measure, we observe a distinction between the two lines, suggesting the existence of dependency between the proximity measures and M&A transactions (the differences in the two lower subplots are visually less distinct because both social and investor proximity measures are discrete and have a large point mass at 0). Next, we appeal to a more rigorous statistical model for further analysis.

Statistical Model

Using statistical terminology, the matching of a pair of firms is a binary outcome: Either they are part of an M&A transaction or they are not. Thus, it could be tempting to use the binary response econometric models such as logistic regression for empirical analysis. However, logistic regression assumes independent observations. In our context, it means inter-firm transactions are independent of each other; whether an M&A transaction occurs between firms i and j is independent of any other transaction(s). This assumption is implausible due to the relational nature of the M&A data. For example, an M&A transaction between firms i and j and that between i and k (which would be two observations in a logistic regression) are correlated since they involve a common party (i.e., firm i). If j had acquired i, then k couldn't have acquired i: one company could not be bought twice. Hence, the key assumption of independent observations, which underlies the binary response econometric models, is clearly violated. So instead of treating the M&A transactions as independent observations, we model all of them together as a *network*.



Note: In (b), we plot geographic distance rather than geographic proximity.

Figure 5. Distributions of Proximity: M&A Sample Versus Random Sample

Exponential random graph models (ERGMs), also known as *p** models, have been developed in statistical network analysis over the past three decades and recently have become perhaps the most important and popular class of statistical models of network structure (for a survey of models in this field, see Goldenberg et al. 2010). As far as we are aware, this modeling framework has not been widely used in the information systems literature thus far, so we briefly introduce it here.[10] We also provide a list of important notations used in this and the following sections in Table A1 in the appendix for reference.

A network is a way to represent relational data in the form of a mathematical *graph*. A graph consists of a set of *nodes* and a set of *edges*, where an edge is a directed or undirected link between a pair of nodes. A network of n nodes can also be mathematically represented by an $n \times n$ adjacency matrix Y, where each element Y_{ij} can be zero or one, with one indicating the existence of the i–j edge and zero meaning otherwise. Self-edges are disallowed so $Y_{ii} = 0\ \forall i$. If edges are undirected (i.e., the i–j edge is not distinguished from the j–i edge), then $Y_{ij} = Y_{ji}\ \forall i, j$ (i.e., Y is a symmetric matrix).

In applications, the nodes in a network are used to represent economic or social entities, and the edges are used to represent certain relations between the entities. In this present research, the nodes and the edges are high-tech companies and the M&A transactions between them respectively, and

[10]The only papers using ERGMs by information systems scholars that we are aware of are Skerlavaj et al. (2010) and Faraj and Johnson (2011).

they together form an M&A network. In terms of the adjacency-matrix representation, we define Y_{ij}=1 if *i* and *j* are part of an M&A transaction and 0 otherwise. With this definition, the resultant M&A network is undirected.[11]

ERGMs treat network graph, or equivalently adjacency matrix *Y*, as a random outcome. For a network of *n* nodes, the set of all possible graphs (denoted $\mathcal{Y}$) is finite. The observed network is one realization of the underlying random graph generation process. For some $y \in \mathcal{Y}$, the probability of it occurring is assumed to be

$$P(Y=y)=\frac{1}{\psi}\exp\left\{\sum_{k=1}^{K}\theta_k z_k(y)\right\} \tag{8}$$

where *K* is the number of network statistics, $z_k(y)$ is the k^{th} network statistic, the θ_k's are parameters, and the denominator Ψ is a normalizing constant.[12] The $z_k(y)$ terms capture certain properties of the network and are assumed to affect the likelihood of its occurring. They are analogous to the independent variables in a regression model. One common example of network statistics is the total number of edges in the network (or a constant multiple of it). $z_k(y)$ can be a function of not only the network graph *y*, but also other exogenous covariates on the nodes. For example, suppose we have a categorical variable on the nodes. Then one such statistic is the number of edges where the two ending nodes belong to the same category. To interpret the parameters θ_k, we can rewrite equation (8) in terms of log-odds of the conditional probability (Goldenberg et al. 2010)

$$\text{logit}\left(P\left(Y_{ij}=1|Y_{-ij}\right)\right)=\sum_{k=1}^{K}\theta_k\Delta z_k \tag{9}$$

where Y_{-ij} is all but the *ij* element in the adjacency matrix. Therefore, the interpretation of θ_k is: If forming the *i*–*j* edge increases z_k by 1 and the other statistics stay constant, then the log-odds of it forming is θ_k.[13, 14]

Specification

Our ERGM specification includes the statistics (z_k's) for degree distribution, selective mixing, and proximity. We iterate them and explain their interpretations in the M&A context in the following paragraphs. In the discussion, we translate the generic terms *nodes* and *edges* into the more specific terms *firms* and *transactions*.

The degree distribution statistics include *t*, the total number of M&A transactions, and d_2, the number of firms that each are a party of at least two different transactions. *t* measures the density of transactions in the M&A network and its coefficient serves a similar role as the constant term in a regression model. In fact, equation (9) implies that the coefficient of *t* is the log-odds of transaction happening if *t* were the only statistic in the equation. Given the sparsity of the M&A network, we expect *t*'s coefficient to be negative. The reason why we also include the d_2 statistic is because it has been demonstrated in the prior research that firms with different relational capabilities (Lorenzoni and Lipparini 1999) participate in significantly different levels of M&A activities. Wang and Zajac (2007) specifically showed that an acquisition is more likely to occur if any of the two parties have prior acquisition experiences. Moreover, we have found in the exploratory data analysis in the "Data" section that the number of M&A transactions in which a firm is a party follows the power-law distribution. Hence we predict a transaction where either of the two parties that has previously engaged in M&A transactions should have a different likelihood than when neither has. The d_2 statistic captures exactly this effect and we expect its coefficient to be positive.

[11] Alternatively, we could define a directed "acquisition network" where the edges are asymmetric. That is, we could distinguish the acquirer and the acquired. For our purposes of assessing the business proximity measure, the distinction is not very important since business proximity is symmetric (and it is also true for the other three proximity measures). In addition, our assumption of undirected M&A network reduces the time needed for computation when we perform the estimations.

[12] $\sum_{y\in\mathcal{Y}}P(Y=y)=1$, so $\psi=\sum_{y\in\mathcal{Y}}\exp\left\{\sum_{k=1}^{K}\theta_k\Delta z_k\right\}$.

[13] It is noteworthy that if the Δz_k's do not depend on Y_{ij} $\forall i, j$, then the edges are independent of each other, and hence the ERGM model reduces to a standard logistic regression where each edge is considered an independent observation.

[14] The above summarizes the basic formulation of ERGMs. Despite its relatively straightforward interpretation and analytic convergence, applications had been limited until just a few years ago due to significant computational burdens. The difficulty lies in evaluating the normalizing constant in equation (8), which involves a sum over a very large sample space even for a moderate *n*. It is not hard to see that the number of possible graphis is $2^{n(n-1)}$ if the network is directed, and the number of possible graphs is $2^{n(n-1)/2}$ if the network is undirected. Recent advances in computing capability and Monte Carlo estimation techniques (e.g., Handcock et al. 208; Snijders 2002) have made possible the significant growth of ERGMs applications in academic fields such as sociology and demography.

Selective mixing captures the matching of firms according to the combination of their *nodal-level* characteristics. In other words, these characteristics are first defined at the individual firm level, and then combined at the pair level and, finally, aggregated to the corresponding network statistics. In the network analysis literature, one widely adopted form of selective mixing is assortative mixing: Social and economic entities tend to form relationships with others that are "similar." We include two groups of statistics that reflect an analogous kind of selective mixing in M&As and they are constructed based on two categorical covariates we have on the firms (i.e., state and industry sector). We expect that a pair of firms belonging to the same category are more likely to match than otherwise. Specifically, statistic h_s^{sta} is the number of transactions between two firms whose headquarters are both located in state s, where s is one of the 50 states plus the District of Columbia; h_c^{cat} is the number of transactions between two firms that belong to the same industry sector c, where c is any of the 19 sectors described in the "Data" section. We also want to point out that these two groups of statistics can serve as alternative operationalizations of geographic and business proximity.

Finally, the statistics of most interest are the four proximity measures that capture the matching process based on *dyadic-level* characteristics. We normalize the four proximity measures to ensure they have the same standard deviation. The four statistics each equal the sum of the corresponding characteristic values over all transactions. We use p_g, p_s, and p_b to denote the sums of geographic proximity, social proximity, investor proximity, and business proximity, respectively. The rationale of including them was discussed earlier. In the benchmark specification, we include a linear term for p_b. We also estimate an additional specification with a quadratic term of p_b to allow for a curvilinear effect of business proximity on matching.

To sum up, our benchmark model specification can be written

$$P(Y=y)=\frac{1}{\psi}\exp\left\{\theta_l t+\theta_{d_2}d_2+\sum_s\theta_s^{sta}h_s^{sta} +\sum_c\theta_c^{cat}h_c^{cat}+\theta_g p_g+\theta_s p_s+\theta_f p_f+\theta_b p_b\right\} \tag{10}$$

and the corresponding conditional form is

$$\begin{aligned}&\operatorname{logit}\left(P\left(Y_{ij}=1|Y_{-ij}\right)\right)\\&=\theta_l\Delta t+\theta_{d_2}\Delta d_2+\sum_s\theta_s^{sta}\Delta h_s^{sta}+\sum_c\theta_c^{cat}\Delta h_c^{cat}+\theta_g\Delta p_g\\&\quad+\theta_s\Delta p_s+\theta_f\Delta p_f+\theta_b\Delta p_b\\&=\theta_l+\theta_{d_2}\Delta d_2+\sum_s\theta_s^{sta}I\left(s_i=s_j=s\right)+\sum_c\theta_c^{cat}I\left(c_i=c_j=c\right)\\&\quad+\theta_g\Delta p_{g,ij}+\theta_s\Delta p_{s,ij}+\theta_f\Delta p_{f,ij}+\theta_b\Delta p_{b,ij}\end{aligned} \tag{11}$$

where $I(\cdot)$ is an indicator function, and, for instance, $I(s_i = s_j = s)$ means companies i and j are in the same state s and $I(c_i = c_j = c)$ means i and j belong to the same sector c.

Results

The final dataset contains a total of 24,382 companies. This seemingly moderate number of nodes is actually huge for estimating network models, since the number of potential edges (in our case, unordered pairs) is close to 300 million. Given our current computational capacity, we cannot handle the whole dataset in one estimation procedure. To carry out the analysis, we decide to randomly select 25% of the whole dataset for estimation and repeatedly do so 100 times. Since the estimation for each subsample is an independent, computationally intensive task, we parallelized the estimation job using the Condor system,[15] which is a big data platform to support high throughput computing. For each of the 100 different samples (6,096 companies each), we estimate the model coefficients by using the Markov Chain Monte Carlo maximum likelihood estimation procedure outlined in Hunter and Handcock (2006).

We summarize the resultant 100 set of coefficients for the degree distribution, selective mixing, and proximity statistics in Tables 2, 3, and 4, respectively. For each statistic, we report the number of samples that yield a coefficient with the expected sign, and the number(s) of samples that yield a coefficient that has the expected sign and is statistically significant at one or more selected confidence level(s). Also, to provide an example, we report the full estimation result for one particular sample in Table A2 in the appendix.

Table 2 reports the coefficients of the degree distribution statistics. Among the 100 samples, all θ_t coefficients are negative and 97 θ_{d2} coefficients are positive. At the 99.0% confidence level, 98 θ_t estimates are significant and 92 θ_{d2} estimates are significant. Hence the results for the two degree distribution statistics are both consistent with our expectations. As discussed, the negativity of θ_t indicates only the overall small probability of an M&A transaction occurring; the positive sign of θ_{d2} means that an M&A transaction of which firms with some M&A experience are involved is more likely to occur.

In panel (a) of Table 3, we find most state-based selective mixing statistics are dropped. This is due the sparsity of M&A transactions during the data collection period: the likelihood that two same-state companies merged in an individual sample is low for most states. Indeed, the states that yield the

[15] http://research.cs.wisc.edu/htcondor/.

Table 2. Degree Distribution Coefficients (100 Samples)

		Number of Samples with Expected Sign	Number of Samples with *p*-value	Median Coefficient Value
θ_t	edge	100 (< 0)	98	-14.7837
θ_{d2}	degree > 2	97 (> 0)	92	3.0064

Table 3. Selective Mixing Coefficients (100 Samples)

	Number of Samples with Coefficients	Number of Samples Coefficient > 0	Number of Samples *p*-value < 1.0%		Number of Samples with Coefficients	Number of Samples Coefficient > 0	Number of Samples *p*-value < 1.0%
Panel A: State							
AK	0	–	–	MT	0	–	–
AL	0	–	–	NC	0	–	–
AR	0	–	–	ND	0	–	–
AZ	0	–	–	NE	0	–	–
CA	100	94	43	NH	5	5	3
CO	7	7	7	NJ	4	4	3
CT	0	–	–	NM	0	–	–
DC	5	5	4	NV	0	–	–
DE	0	–	–	NY	61	61	22
FL	0	–	–	OH	0	–	–
GA	7	7	6	OK	0	–	–
HI	0	–	–	OR	0	–	–
IA	0	–	–	PA	0	–	–
ID	0	–	–	RI	0	–	–
IL	5	5	5	SC	0	–	–
IN	0	–	–	SD	0	–	–
KS	0	–	–	TN	0	–	–
KY	0	–	–	TX	19	19	13
LA	0	–	–	UT	0	–	–
MA	28	28	16	VA	0	–	–
MD	6	6	5	VT	0	–	–
ME	0	–	–	WA	11	11	6
MI	0	–	–	WI	0	–	–
MN	0	–	–	WV	0	–	–
MO	0	–	–	WY	0	–	–
MS	0	–	–				
Panel B: Category							
advertising	26	25	7	mobile	28	26	11
biotech	38	37	5	net hosting	7	6	6
cleantech	11	11	6	other	0	–	–
consulting	11	10	3	pub rel	8	8	8
ecommerce	13	13	3	search	0	–	–
education	0	–	–	security	0	–	–
enterprise	22	22	20	semiconductor	15	15	5
games video	26	25	11	software	87	78	37
hardware	32	31	25	web	76	65	21
legal	0	–	–				

Table 4. Proximity Coefficients (100 Samples)

		Number of Samples with Coefficient > 0	Number of Samples with *p*-value < 5.0%	Number of Samples with *p*-value < 1.0%	Number of Samples with *p*-value < 0.1%	Median Estimate
θ_g	Geographic	46	8	5	3	-0.0173
θ_s	Social	79	73	70	69	0.1460
θ_t	Investor	62	52	51	46	0.0689
θ_b	Business	100	92	86	79	0.5315

Table 5 Proximity Coefficients (100 Samples): Equation (10) plus $\theta_{b2}p_{b2}$

		Number of Samples with Expected Sign	Number of Samples with *p*-value < 5.0%	Number of Samples with *p*-value < 10%	Number of Samples with *p*-value < 0.1%
θ_g	Geographic	47 (> 0)	6	4	2
θ_s	Social	85 (> 0)	77	77	73
θ_t	Investor	67 (> 0)	56	52	50
θ_b	Business	100 (> 0)	86	76	61
θ_{b2}	Business²	86 (< 0)	42	28	13

most coefficients, namely CA, NY, and MA, are where well-known high-tech hubs are located. In panel (b) of Table 3, we observe that for almost all category-based selective mixing statistics, an overwhelmingly large proportion of the coefficient estimates are positive, but it turns out their statistical significance, when using the 99.0% confidence level, is not strongly supported. One possible explanation of their statistical insignificance is the inclusion of our business proximity measure. As mentioned, the selective mixing statistics based on industry sector can also be thought of as alternative, but coarser, operationalizations of business proximity. Therefore, when including both the selective mixing statistics and our business proximity measure in the ERGM specification, the effect of the selective mixing statistics is superseded by the effect of the more refined proximity measure, causing the model to produce insignificant coefficients for the selective mixing statistics. To test the validity of this explanation, we also estimate another ERGM specification, which excludes the business proximity measures and for which we report the corresponding results for the selective mixing coefficients in Table A3 in the appendix. Comparing the last columns of Tables 3 and A3, we find that when using the specification without proposed business proximity, a much higher proportion of the samples produces statistically significant (at the 1.0% significance level) estimates for the selective mixing coefficients. This is thus supporting evidence for the superiority of the proximity measures we use: They are correlated with the alternative, coarser measures, but statistically more powerful in explaining the matching in M&As.

In Table 4 we report the estimation results for the four proximity measures. First and foremost, the new business proximity measure is found to be strongly associated with the matching likelihood: All of the samples produce positive coefficients and among them 79 estimates are significant at the 99.9% confidence level. Furthermore, when comparing the proximity measures across the rows, we observe three among the four proximity measures (except θ_g geographic) are positively associated with the likelihood of matching in M&As, and, in particular, our newly developed business proximity measure also outperforms the other three in terms of statistical significance. Moreover, since we normalize the proximity measures, we can evaluate their economic significance by comparing the magnitude of the coefficients. Using the median estimate from the 100 samples (last column of Table 4), we find that the business proximity measure has the largest effect on the matching likelihood: A 1 standard deviation increase in business proximity has the same effect as a 3.64 standard deviation increase in social proximity, or a 6.89 standard deviation increase in investor proximity. These results thus support the value of business proximity in modeling M&As. Interestingly, in our dataset, the geographic proximity appears to play an insignificant role in identifying high-tech firms' matching in M&As.

The estimation result of equation (10) shows business proximity is positively associated with the M&A matching likelihood. However, a linear structure might not best capture the true relationship between business proximity and M&A matching since the economic benefits of merging two firms'

businesses may result from not only their similarity but also their complementarity (e.g., Chung et al. 2000; Sears and Hoetker 2013). The value of M&A could decrease in cases where two firms' businesses are too similar but lack complementarity, so little value of synergy can be achieved through merger. We test this hypothesis by estimating a specification that includes a squared term of business proximity, $\theta_{b2}p_{b2} = \theta_{b2}\sum p^2_{b,ij}$ and that is otherwise the same as equation (10). We expect θ_{b2} to be negative and θ_b to be still positive. The estimation results on the proximity measures (of the 100 samples) are reported in Table 5. We do observe that for a large number of the samples, business proximity is estimated to have a curvilinear effect on the M&A matching likelihood. Specifically, for 86 out of the 100 samples, the coefficient of the squared term is negative and that of the linear term is positive, suggesting the matching likelihood first increases with business proximity and then decreases after a certain point. This evidence is thus consistent with our expectation. Meanwhile, we note that the statistical significance of the squared term is not as strong as that for the linear term.

Scaling up to Big Data: A System Prototype for Navigating the Networked Startup World

During the recent boom of the high-tech industry, the media are often full of reports about high-profile M&As involving startups. It is well known that M&As are an important alternative to IPOs as an exit option for high-tech entrepreneurs and early investors. Meanwhile, industry giants spend tens of billions of dollars each year in acquiring smaller firms for market entrance, strategic intellectual property, and talented employees.[16] Venture capitalists also arrange mergers between their partially owned startups in order to consolidate resources and reduce competitive pressure.[17] The fierce competition in both demand and supply instantaneously creates the problem of matching between acquirers and targets, since the value (or disvalue) of an M&A critically depends on the synergy of the companies' products, technologies, and markets. Broadly, the challenge lies in the search for startups. While almost everyone knows who the top competitors are in a particular space, it is a difficult and time-consuming task to find the small companies in the vast startup universe with the right products or technology. The problem can become increasingly challenging over time given the speed of technological innovation. Solving this search problem will be beneficial not only for M&A executives, but also for entrepreneurs to position their products and identify competitors, for venture capitalists to monitor niche markets, and for high-tech analysts to examine the industry trend. Observers have noted data analytics can complement executives' industry knowledge in alleviating many of the problems, and transform the way M&A matching and startup search have been done; it is reported that many large M&A players have already been investing heavily in analytics for identifying the win–win matches by rendering the decision-making processes more "data-driven."[18]

Along these lines, our empirical analysis indicates the potential practical value of the proposed business proximity measure as an important metric in the analytics of M&A matching and a search tool for navigating the networked startup world. To show the practical application in a concrete way, we build a prototype for a cloud-based information system that allows entrepreneurs, managers, and analysts to explore the competitive landscape of the U.S. high-tech industry (Whinston and Geng 2004). By incorporating business proximity and making it explicitly available to the users in the search and navigation tools, the platform expedites the process of startup search and competition analysis as well as facilitates efficient new niche-market discovery. Built on the latest big data and cloud technologies, the system largely consists of two components as shown in Figure 6: The back end collects raw data from the data sources, integrates and cleans the data, computes business proximity, and stores the processed data in local databases. The front end is a web application that enables users to explore the data stored in a cloud-based database.

Back-End System

The back-end system comprises two modules and two databases. The first module is the data collector written in Python to retrieve data from our data sources, including CrunchBase. The collector runs periodically to ensure our data is up-to-date. The raw data is stored in a MongoDB[19] database, which is a document-oriented, NoSQL database that stores records in JSON format. The reason why we do not use a relational

[16]See "Internet Mergers and Takeovers: Platforms upon Platforms," *The Economist*, May 25, 2013.

[17]An example is the acquisition of Summarize by Twitter in 2008. See "Finding a Perfect Match," *Twitter Blog*, https://blog.twitter.com/2008/finding-perfect-match and Nick Bilton's 2013 book, *Hatching Twitter: A True Story of Money, Power, Friendship, and Betrayal.*

[18]See "Google Ventures Stresses Science of Deal, Not Art of the Deal," *New York Times*, June 23, 2013, and "One of the Richest Men in the World Is Backing a Startup that Ranks Wall Street's Hedge Funds," *Business Insider*, http://read.bi/1KqhHzr.

[19]https://www.mongodb.org.

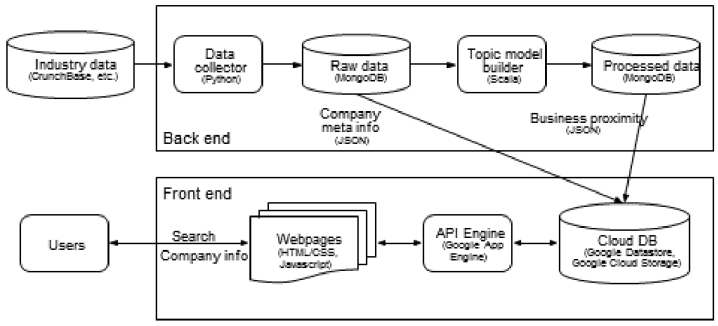


Figure 6. Prototype Architecture and Components

database is that the structure of the company data may change over time, so the traditional relational database, which requires a predefined schema, is not the best technology for our system. Another feature of MongoDB is that it supports scalability: As the data size grows, load balancing can be performed using a shrading mechanism. This is a basis for the cloud-based information system.

The second module, the topic model builder, constructs and estimates topic models using the textual company descriptions extracted from the raw data in MongoDB. To run the LDA topic modeling algorithm, we use a Scala implementation in Stanford Topic Model Toolkit.[20] The topic model builder produces two sets of results: First, underlying business topics of the whole industry are generated, where each topic is essentially a set of related keywords that represent the topic. Second, each company's profile is transformed into a topic vector, which is stored in the database of processed data in MongoDB.

We then compute business proximity to identify the top N nearest neighbors from each firm. A naive, brute-force approach that calculates the business proximity values for all pairs of companies can be used to find the nearest neighbors. However, as we continuously collect data and the dataset grows, the number of company pairs increases exponentially to a point that the exhaustive computation is impractical for the real-world system. Hence, we propose an algorithm that reduces the required computation while providing a reasonable approximation in finding nearest neighbors. The intuition behind the algorithm is that a pair of companies is likely to have a high proximity value only if they share high weights on some common topics in their topic distributions. Hence, we maintain a bucket list for each topic that keeps track of the companies with a high weight on that specific topic. Then we only compute the business proximity values for company pairs that co-occur in at least one of the bucket lists, because those pairs that do not fall into any of the bucket lists are unlikely to be very close to each other. The pseudocode is given in Algorithm 1.

To measure the speed of business proximity computation and the accuracy of the nearest-neighbor identification, we run experiments using the dataset described in the "Data" section. The results are reported in Figure 7. In terms of the computation speed, we count the number of business proximity values calculated. We use this metric instead of the actual computation time to avoid potential environmental biases. The brute-force algorithm, which computes all pairwise proximity values, requires 341 million calculations. In the meantime, our algorithm with threshold 0.00 only needs 123 million, which is 36% of the naive approach. As we increase the threshold to 0.30, only 3% of calculations are needed. Faster computation comes with a modest cost on accuracy. We compare the N nearest neighbors identified by the algorithm with different thresholds and vary N to be 10, 20, 30, 50, and 100. As expected, the algorithm provides accurate results for closest neighbors, where the performance degrades gracefully to the not-so-near neighbors. We want to note that the algorithm with threshold 0.00 provides 100% accurate neighborhood sets comparing to the brute-force algorithm. Even for the case of threshold 0.30, the algorithm gives a 92.5% accuracy in identifying 50 nearest neighbors.

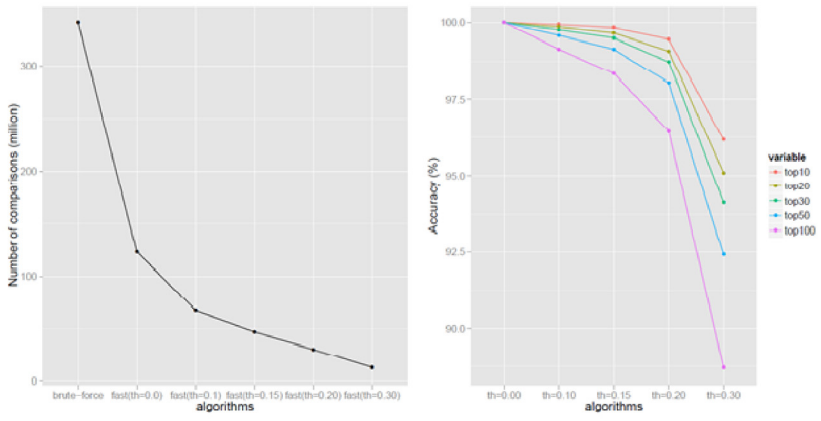
Front-End System

The front end is a cloud-based web application, which is available at http://diamond.mccombs.utexas.edu/bizprox, to let users explore various company information with the proposed business proximity. Figure 8 shows the screenshots of

[20]http://www-nlp.stanford.edu/software/tmt/tmt-0.4/.

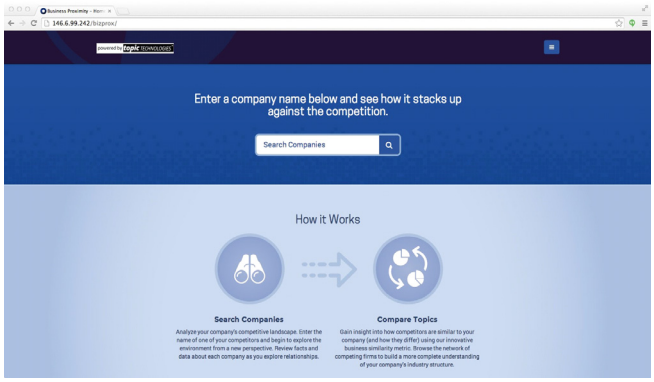
input: set of companies C, companies' topic distributions T, number of topics K, and threshold θ
output: N nearest neighbors for each company
for *each topic* $k \in K$ **do**
    $B_k \leftarrow \theta$
end
for *each company* $c \in C$ **do**
    for *each topic* $k \in K$ **do**
        if $T[c][k] >= \theta$ **then**
            $B_k \leftarrow B_k \cup c$
        end
    end
end
for *each company* $c \in C$ **do**
    $N\ set \leftarrow \emptyset$
    for *each topic* $k \in K$ **do**
        if $T[c][k] >= \theta$ **then**
            $N\ set \leftarrow N\ set \cup B_k$
        end
    end
    for *each company* $c' \in N\ set$ **do**
        $bizprox[c'] \leftarrow cosine\_similarity(T[c], T[c'])$
    end
    Find N nearest neighbors by sorting *bizprox* list
end

Algorithm 1: Faster Nearest-Neighbor Computation

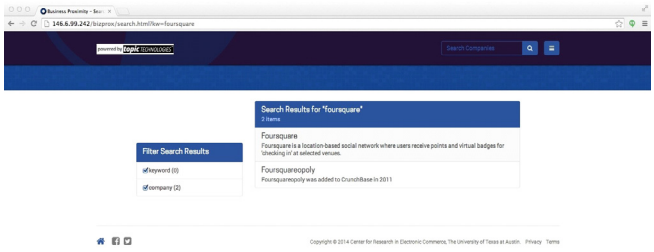


(a) Calculation Speed **(b) Accuracy of Nearest Neighbors**

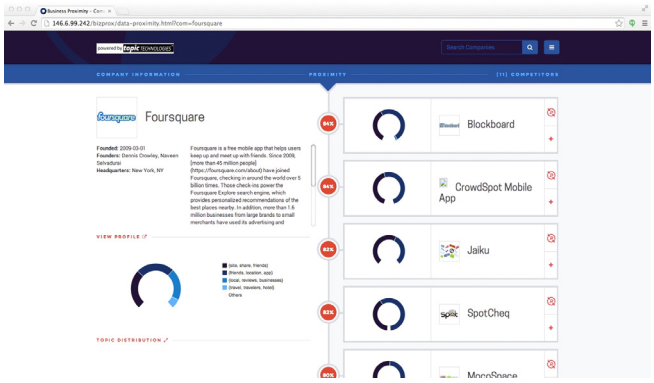
Figure 7. Performance Measures of Algorithm 1



(a) Search Companies and Topics of Interest



(b) Search Results



(c) Focal Company with its Competitors Based on Business Proximity

Figure 8. Prototype Front End: User Interface Screenshots

the user interface. Given a keyword from the user, the search results show the topics and companies associated to the keyword. By selecting topics, the user can interpret the topic with 20 (additional) relevant keywords and the significance of each. If a company is selected from the search results, the interface provides (1) the basic information about the company along with the topic distribution, and (2) a list of nearest neighbors to the focal company. The basic information of a company includes the founding date, founders, headquarters, and a short business description. With the topic distribution, users can recognize various business aspects of the company. The nearest neighbors are computed using Algorithm 1 and are sorted by the business proximity.

From the system architecture perspective, the front end is a cloud-based system leveraging platform-as-a-service (PaaS). The static webpages in HTML/CSS are hosted by our local Apache Web Server. The server interacts with the various

user inputs such as keyword searches and page navigations. Each webpage is instrumented with Google Analytics[21] so that web analytics are performed to understand user engagement and potentially optimize the service. An API Engine, deployed in Google App Engine,[22] receives queries from the HTML pages and returns relevant data from the cloud database. The cloud database consists of two components: First, the dynamic data is managed in Google Cloud Datastore,[23] a cloud-based NoSQL database system; second, the static data is stored in Google Cloud Storage,[24] which provides a cost-effective content distribution service for static information. The cloud-based approach gives two main benefits: scalability (e.g., the system scales automatically according to user demand and data size) and availability (e.g., almost no downtime due to replication).

Discussion and Conclusion

The advent of the digital economy is creating a business environment that is characterized by the unprecedented complexity of technology and connectedness between firms and people. With the goal of reducing the difficulty to understand and depict the business landscape, in this paper we set out to develop a general, data-analytic framework for quantifying firms' positions in the spaces of product, market, and technology and for measuring firms' dyadic business proximity. Using a unique dataset of the U.S. high-tech industry as an example, we detailed the procedure that uses topic models to analyze the publicly available, textual descriptions of company business and constructs proximity according to the structured results. We then validated the new measure by relating it to the simple category-based classification and analyzing its statistical relationships with firm interactions including M&A, investment, and job mobility. In a more rigorous statistical analysis, we also demonstrated the new measure's usefulness in modeling matching of M&As, where we constructed a network of high-tech companies and documented empirical evidence on the nuanced relationship between matching and business proximity. We found the statistical significance of business proximity in explaining M&A matching to be the strongest compared with geographic, social, and investor proximities. Moreover, to show the practical value of the proposed data-analytic framework, we deployed various big data and analytics technologies to build a prototype of a cloud-based information system for industry intelligence.

This research sheds light on the value of leveraging data science techniques in the development of novel measures (Einav and Levin 2013) for large-scale business analytics. Our data-driven, analytics-based approach requires no expert preprocessing, provides finer granularity (compared with the SIC- or NAICS-based methods), is more comprehensive on quantifying firms' positions in the spaces of product, market, and technology (compared with the patent- or customer-based methods), and can be better automated and scaled to big data (compared with all previous methods). When built into an automated system as in the previous section, the method is also more responsive in capturing industry trends than any human-annotation-based approach. Substantively, the comprehensive, granular business proximity measure is an enabler in the M&A application to show the nuanced relationship between the transaction likelihood and the firms' business similarity and complementarity. The result manifests the fact that economically meaningful information can be extracted from unstructured data through careful analysis and large-scale computation. Thus, our methodology greatly complements the toolkit for measuring business proximity. It is especially useful when researchers or analysts are studying either an already narrowly focused industry or a highly dynamic industry or when the firms under study are small and privately held (e.g., startups) so industry classification is largely unavailable. Meanwhile, we wish to stress that our measure is not intended as a one-stop replacement for all existing methods. Rather, researchers should evaluate which method fits the research purpose best and there are scenarios where some previous method is sufficient. For instance, if the study focuses on R&D, then the patent-based method may serve the research purpose well; or if the research question is at a relatively macro level, only firms' broad industry membership is important, and all firms' SIC or NAICS codes are available, the researcher should not be hesitant to use the SIC- or NAICS-based methods.

More broadly, the data-analytic framework used in the study presents a general approach for understanding industry structure and it also demonstrates the potential transformation big data analytics can bring into both industry intelligence practice and strategy and industrial organization research. For analytics-minded managers, firms' relatedness in business is a very important metric for identifying potential partners, competitors, and alliance or acquisition targets. The saying in management goes, "if you cannot measure it, you cannot manage it." As shown in our study, the proposed proximity measure provides finer granularity, and is proved to be effective in high-tech M&A analytics. More importantly, as a

[21] http://www.google.com/analytics/.

[22] https://developers.google.com/appengine/.

[23] https://developers.google.com/datastore/.

[24] https://cloud.google.com/products/cloud-storage/.

general approach to organize unstructured data for industry intelligence, the usefulness of the proposed framework is not limited to measuring proximity and analyzing M&As. Rather, as argued and demonstrated in the previous section, it provides a handy leverage for entrepreneurs, venture capitalists, and analysts to navigate the constantly changing landscape of the networked business environment, which is much needed in light of the rapid evolution of technology and increasing complexity of the digital economy. Our prototype can be the first step in building a business intelligence platform to fully realize the framework's practical potential. In response to the transformation, even for outside the domain of industry intelligence, organizations need to invest in IT infrastructure and capability to better organize and analyze unstructured data, as the ability of distilling value from unstructured data will be an important competitive advantage in the digital economy. Our prototype is also an example of organizing unstructured data and integrating the state-of-the-art storage and computation technologies to build a decision support system. For business and economics scholars, our method can perhaps be adapted and serve as an alternative approach of defining market boundary or identifying industry rivals, which is a crucial step in the empirical research of industrial organization. Additionally, future research can explore the possibility of combining topic modeling results and clustering algorithms to build an industry hierarchy, which could be a data-driven alternative to the expert-labeled systems that are currently in use. A data-driven approach is especially desirable for industries such as high-tech because the underlying technology is rapidly changing and the manually labeled industry classification system can be stale.

This research also advances the understanding and analysis of M&As. We documented systematic evidence on the relationship between M&A matching and firm proximity in the high-tech industry, which complements the previous empirical M&A literature that primarily focused on larger, public corporations (Betton et al. 2008). The proposed new measure also enabled us to test the non-monotone relationship between business proximity and M&A matching. More importantly, we constructed a network structure using firm proximity measured in four different dimensions and adopted the statistical modeling framework of ERGMs to accommodate the relational nature of the matching data. The network/graph approach has been fruitfully applied to analyzing a variety of economic exchanges and markets (as surveyed in Easley and Kleinberg (2010) and Jackson (2010)). However, whereas the literature is abundant with studies on how networks affect the interaction and performance of firms, research using rigorous statistical methods to analyze the structure of inter-firm networks is relatively underdeveloped. To our knowledge, the M&A application in this study is the first to use a statistical network model to analyze relational transactions among companies. We believe statistical network models are currently underutilized by management scholars in their empirical research on interorganizational linkage despite the fact that relational data is actually not uncommon in the studies of many important questions. For example, strategic alliances, investments, and patent license agreements among companies can all be visualized and carefully analyzed as graphs/networks. We predict that with the growing availability of network datasets and ongoing development of large-scale computing technologies, the value of statistical network models in management research will be increasingly recognized.

In closing, we wish to point out some additional caveats and limitations of the research. First, since SIC- or NAICS-based industry classification or patent data is unavailable for most companies in CrunchBase, we could not directly compare the proposed business proximity measure with that based on industry hierarchy (Wang and Zajac 2007) or the measure based on patent citation (Stuart 1998) in terms of their explanatory power for M&A matching. Although this is less crucial for this paper, since our goal is not to search for the best empirical model for M&As, it could be an interesting research project to find a suitable dataset where all the new and traditional measures could be operationalized and compared directly. Second, for our data-analytic approach, the number of topics in LDA is a free parameter for users to choose. When performing topic modeling on the CrunchBase descriptions, we selected a finite set of values for this parameter, which is sufficient for our purpose of illustrating the general methodology. Nevertheless, from a practical point of view, it is worth investigating whether an "optimal" number of topics exists, and if so, how it can be determined (for example, by leveraging the newly developed method in Lancichinetti et al. (2015)). Third, in the machine-learning literature, there are several extensions to the LDA algorithm (e.g., Inoyue et al. 2014; Teh et al. 2004). Future research could investigate how these extensions could benefit understanding company businesses through text analysis. Fourth, some company-level characteristics that are commonly used in the finance literature (notably, company size and revenue) are unavailable in our dataset, which inevitably limited our ability to extend our empirical application on M&A matching. For instance, had we observed company size, we would be able to study the moderating effect of companies' size on the relationship between business proximity and the matching likelihood. Another related point is that because the vast majority of companies in our dataset are private, there is no publicly available financial statement information (for example, SEC filings) for them. However, public filings are probably more comprehensive and less error prone than information provided by third-party data vendors. Thus for future researchers who focus on public firms, they could use

the firms' public filings to SEC in the proposed data-analytic framework. Finally, the model we employed in the empirical analysis, ERGM, is a static network model. The area of statistical network analysis is developing fast. To deepen our understanding about the dependency structure of M&A transactions, future research could leverage certain newer dynamic network models to examine the evolution of the M&A network.

Acknowledgments

We thank the special issue editors and two anonymous reviewers for their guidance and constructive suggestions during the review process. We thank Stephen Whinston and Ying-Yu Chen for their work on developing the interface of our system prototype. We also thank Mark Varga for his help on converting our LaTeX manuscript to Microsoft Word format.

References

Adomavicius, G., and Tuzhilin, A. 2005. "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions," *IEEE Transactions on Knowledge and Data Engineering* (17:6), pp. 734-749.

Ahuja, G., and Katila, R. 2001. "Technological Acquisitions and the Innovation Performance of Acquiring Firms: A Longitudinal Study," *Strategic Management Journal* (22:3), pp. 197-220.

Amit, R., Glosten, L., and Muller, E. 1990. "Entrepreneurial Ability, Venture Investments, and Risk Sharing,"*Management Science* (36:10), pp. 1233-1246.

Betton, S., Eckbo, B. E., and Thorburn, K. S. 2008. "Corporate Takeovers," Chapter 15 in *Handbook of Corporate Finance: Empirical Corporate Finance*, Volume 2 (1st ed.), B. E. Eckbo (ed.), Amsterdam: Elsevier/North-Holland, pp. 291-430.

Blei, D. M. 2012. "Introduction to Probabilistic Topic Models," *Communications of the ACM* (55:4), pp. 77-84.

Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. "Latent Dirichlet Allocation," *Journal of Machine Learning Research* (3), pp. 993-1022.

Chakrabarti, A., and Mitchell, W. 2013. "The Persistent Effect of Geographic Distance in Acquisition Target Selection," *Organization Science* (24:6), pp. 1805-1826.

Chen, H., Chiang, R. H. L., and Storey, V. C. 2012. "Business Intelligence and Analytics: From Big Data to Big Impact," *MIS Quarterly* (36:4), pp. 1165-1188.

Chiang, R. H. L., Goes, P., and Stohr, E. A. 2012. "Business Intelligence and Analytics Education and Program Development: A Unique Opportunity for the Information Systems Discipline," *ACM Transactions on Management Information Systems* (3:3), pp. 12:1–12:13.

Chung, S., Singh, H., and Lee, K. 2000. "Complementarity, Status Similarity and Social Capital as Drivers of Alliance Formation," *Strategic Management Journal* (21:1), pp. 1-22.

Cohen, L., Frazzini, A., and Malloy, C. J. 2008. "The Small World of Investing: Board Connections and Mutual Fund Returns," *Journal of Political Economy* (116:5), pp. 951-979.

Easley, D., and Kleinberg, J. 2010. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World* New York: Cambridge University Press.

Einav, L., and Levin, J. D. 2013. "The Data Evolution and Economic Analysis," NBER Working Paper 19035.

Erel, I., Liao, R. C., and Weisbach, M. S. 2012. "Determinants of Cross-Border Mergers and Acquisitions," *Journal of Finance* (67:3), pp. 1045-1082.

Fallick. B., Fleischman C. A., and Rebitzer, J. B. 2006. "Job-Hopping in Silicon Valley: Some Evidence Concerning the Microfoundations of a High-Technology Cluster," *The Review of Economics and Statistics* (88:3), pp. 472-481.

Faraj, S., and Johnson, S. L. 2011. "Network Exchange Patterns in Online Communities," *Organization Science* (22:6), pp. 1464-1480.

Ghose, A., Ipeirotis, P. G., and Li, B. 2012. "Designing Ranking Systems for Hotels on Travel Search Engines by Mining User-Generated and Crowd-Sourced Content," *Marketing Science* (31:3), pp. 493-520.

Goldenberg, A., Zheng, A. X., Fienberg, S. E., and Airoldi, E. M. 2010. "A Survey of Statistical Network Models," *Foundations and Trends in Machine Learning* (2:2), pp. 129-233.

Gompers, P. A. 1995. "Optimal Investment, Monitoring, and the Staging of Venture Capital," *Journal of Finance* (50:5), pp. 1461-1489.

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., and Morris, M. 2008. "STATNET: Software Tools for the Representation, Visualization, Analysis and Simulation of Network Data," *Journal of Statistical Software* (24), pp. 24, 1-11.

Hochberg, Y., Ljungqvist, A., and Lu, Y. 2007. "Whom You Know Matters: Venture Capital Networks and Investment Performance," *Journal of Finance* (62:1), pp. 251-301.

Hunter, D. R., and Handcock, M. S. 2006. "Inference in Curved Exponential Family Models for Networks," *Journal of Computational and Graphical Statistics* (15:3), pp. 565-583.

Inouye, D., Ravikumar, P., and Dhillon, I. 2014. "Admixture of Poisson MRFs: A Topic Model with Word Dependencies," in *Proceedings of 31st International Conference on Machine Learning*, Beijing, June 21-26, pp. 683-691.

Jackson, M. O. 2010. *Social and Economic Networks*, Princeton, NJ: Princeton University Press.

Lancichinetti, A., Sirer, M. I., Wang, J. X., Acuna, D., Körding, K., and Nunes Amaral, L. A. 2015. "High-Reproducibility and High-Accuracy Method for Automated Topic Classification," *Physical Review X* (5:1), pp. 011007.

Lorenzoni, G., and Lipparini, A. 1999. "The Leveraging of Interfirm Relationships as a Distinctive Organizational Capability: A Longitudinal Study," *Strategic Management Journal* (20:4), pp. 317-338.

Mikkelson, W. H., and Ruback, R. S. 1985. "An Empirical Analysis of The Interfirm Equity Investment Process," *Journal of Financial Economics* (14:4), pp. 523-553.

Mitsuhashi, H., and Greve, H. R. 2009. "A Matching Theory of Alliance Formation and Organizational Success: Complementarity and Compatibility," *Academy of Management Journal* (52:5), pp. 975-995.

Moscarini, G., and Thomsson, K. 2007. "Occupational and Job Mobility in the US," *Scandinavian Journal of Economics* (109:4), pp. 807-836.

Mowery, D. C., Oxley, J. E., and Silverman, B. S. 1998. "Technological Overlap and Interfirm Cooperation: Implications for The Resource-Based View of The Firm," *Research Policy* (27:5), pp. 507-523.

Rhodes-Kropf, M., and Robinson, D. T. 2008. "The Market for Mergers and the Boundaries of the firm," *Journal of Finance* (63:3), pp. 1161-1211.

Sears, J., and Hoetker, G. 2014. "Technological Overlap, Technological Capabilities, and Resource Recombination in Technological Acquisitions," *Strategic Management Journal* (35:1), pp. 48-67.

Shi, Z., Rui, H., and Whinston, A. B. 2014. "Content Sharing in a Social Broadcasting Environment: Evidence from Twitter," *MIS Quarterly* (38:1), pp. 123-142.

Shmueli, G., and Koppius, O. R. 2011. "Predictive Analytics in Information Systems Research," *MIS Quarterly* (35:3), pp. 553-572.

Skerlavaj, M., Dimovski, V., and Desouza, K. C. 2010. "Patterns and Structures of Intra-Organizational Learning Networks Within a Knowledge-Intensive Organization," *Journal of Information Technology* (25:2), pp. 189-204.

Snijders, T. A. B. 2002. "Markov Chain Monte Carlo Estimation of Exponential Random Graph Models," *Journal of Social Structure* (3:2), pp. 1-40.

Stuart, T. E. 1998. "Network Positions and Propensities to Collaborate: An Investigation of Strategic Alliance Formation in a High-Technology Industry," *Administrative Science Quarterly* (43:3), pp. 668-698.

Stuart, T. E., and Yim, S. 2010. "Board Interlocks and the Propensity to Be Targeted in Private Equity Transactions," *Journal of Financial Economics* (97:1), pp. 174-189.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. 2006. "Hierarchical Dirichlet Processes," *Journal of the American Statistical Association* (101), pp. 1566-1581.

Wang, L., and Zajac, E. J. 2007. "Alliance or Acquisition? A Dyadic Perspective on Interfirm Resource Combinations," *Strategic Management Journal* (28:13), pp. 1291-1317.

Whinston, A. B., and Geng, X. 2004. "Operationalizing The Essential Role of the Information Technology Artifact in Information Systems Research: Gray Area, Pitfalls, and The Importance of Strategic Ambiguity," *MIS Quarterly* (28:2), pp. 149-159.

Xu, L, Duan, J. A., and Whinston, A. B. 2014. "Path to Purchase: A Mutually Exciting Point Process Model for Online Advertising and Conversion," *Management Science* (60:6), pp. 1392-1412.

About the Authors

Zhan (Michael) Shi is an assistant professor of Information Systems at the W. P. Carey School of Business, Arizona State University. Michael's research is at the interface of economics and computation, with applications in social media, online markets, and innovation. A part of this area is now commonly known as data analytics/science. His recent research has been published in top academic journals and conference proceedings including *Journal of Management Information Systems*, *MIS Quarterly*, and the ACM Conference on Economics and Computation.

Gene Moo Lee is an assistant professor of Information Systems at the University of Texas at Arlington. His research interests are in large-scale data analytics with the applications on mobile ecosystems, social network analysis, and Internet security. His works have been published in top conference proceedings, including ACM SIGCOMM Internet Measurement Conference (IMC), IEEE INFOCOM, IEEE SECON, ACM Conference on Economics and Computation (EC), and Workshop on Economics of Information Security (WEIS). He also has extensive industry experience at Samsung Electronics, AT&T Labs, Intel, and Goldman Sachs. He holds 10 patents in mobile technology.

Andrew B. Whinston is the Hugh Cullen Chair Professor in the Information, Risk, and Operation Management Department at the McCombs School of Business at the University of Texas at Austin. He is also the director at the Center for Research in Electronic Commerce. He received his Ph.D. in Economics from Carnegie Mellon University. His recent papers have appeared in *Information Systems Research, Journal of Management Information Systems, MIS Quarterly, Management Science, Marketing Science, Journal of Marketing*, and *Journal of Economic Theory*. He has published over 400 articles in refereed journals, 27 books, and 62 book chapters. In 2005, he received the Leo Award from the Association for Information Systems for his long-term research contribution to the information systems field. In 2009, he was named the Distinguished Fellow by the INFORMS Information Systems Society in recognition of his outstanding intellectual contributions to the information systems discipline. His Erdös number is 2.

MIS Quarterly

BIG DATA & ANALYTICS IN NETWORKED BUSINESS

TOWARD A BETTER MEASURE OF BUSINESS PROXIMITY: TOPIC MODELING FOR INDUSTRY INTELLIGENCE

Zhan (Michael) Shi
Department of Information Systems, W. P. Carey School of Business, Arizona State University,
Tempe, AZ 85287-4606 U.S.A. {zmshi@asu.edu}

Gene Moo Lee
Department of Information Systems and Operations Management, College of Business, The University of Texas at Arlington,
Arlington, TX 76019 U.S.A. {gene.lee@uta.edu}

Andrew B. Whinston
Department of Information, Risk, and Operations Management, McCombs School of Business, The University of Texas at Austin,
Austin, TX 78712 U.S.A. {abw@uts.cc.utexas.edu}

Appendix

Additional Tables

Table A. ERGM Notations

Network Graph	
Y, Y_{ij}	a random network graph matrix, its *i,j* element
Y_{-ij}	all elements except *i,j*
y	the set of all possible graphs for a fixed set of notes y, y_{ij}
y, y_{ij}	a realization of the random network graph and its *i,j* element
$z_k(y)$	a statistic of network graph y
Network Statistics	
t	total number of edges
d_2	number of nodes which have at least 2 edges
h_s^{sta}	number of edges within state *s*
h_c^{cat}	number of edges within category *c*
p_g	sum of geographic proximity over all edges
p_s	sum of social proximity over all edges
p_f	sum of investor proximity over all edges
p_b	sum of business proximity over all edges
Nodal Characteristics:	
s_i	state where *i*'s headquarters is located
c_i	category to which *i* belongs
Dyadic Characteristics	
$p_{g,ij}$	geographic proximity of *i* and *j*
$p_{s,ij}$	social proximity of *i* and *j*
$p_{f,ij}$	investor proximity of *i* and *j*
$p_{b,ij}$	business proximity of *i* and *j*

Table A2. Model Coefficients from Sample 1

	Coeff.	S.E.	*p*-value		Coeff.	S.E.	*p*-value
Geographic	-0.2699	0.3440	0.4326	NY	–	–	–
Social	0.0532	0.0108	0.0000	OH	–	–	–
Investor	0.0270	0.0522	0.6049	OK	–	–	–
Business	0.4635	0.1378	0.0008	OR	–	–	–
Edges	-12.5625	3.7908	0.0009	PA	–	–	–
Degree > 2	2.4820	0.6438	0.0001	RI	–	–	–
				SC	–	–	–
				SD	–	–	–
State				TN	–	–	–
AL	–	–	–	TX	–	–	–
AR	–	–	–	UT	–	–	–
AZ	–	–	–	VA	–	–	–
CA	2.3899	0.8178	0.0035	VT	–	–	–
CO	–	–	–	WA	–	–	–
CT	–	–	–	WI	–	–	–
DC	–	–	–	WV	–	–	–
DE	–	–	–	WY	–	–	–
FL	–	–	–				
GA	–	–	–				
HI	–	–	–	**Category**			
IA	–	–	–	advertising	–	–	–
ID	–	–	–	biotech	–	–	–
IL	–	–	–	cleantech	–	–	–
IN	–	–	–	consulting	–	–	–
KS	–	–	–	ecommerce	–	–	–
KY	–	–	–	education	–	–	–
LA	–	–	–	enterprise	2.9201	0.8882	0.0010
MA	4.6361	1.1201	0.0000	games video	3.0284	1.0953	0.0057
MD	–	–	–	hardware	3.7045	1.7912	0.0386
MN	–	–	–	legal	–	–	–
MO	–	–	–	mobile	1.8611	1.2047	0.1223
MS	–	–	–	network hosting	–	–	–
MT	–	–	–	other	–	–	–
NC	–	–	–	public relations	–	–	–
NE	–	–	–	search	–	–	–
NH	9.7899	1.5931	0.0000	security	–	–	–
NJ	5.6899	1.6428	0.0005	semiconductor	–	–	–
NM	–	–	–	software	–	–	–
NV	–	–	–	web	-0.9020	2.1375	0.6721

Table A3. Category-Based Selective Mixing Coefficients (100 Samples): Equation (10) Excluding $\theta_b p_b$

	Number of Samples with Coefficient	Number of Samples Coefficient > 0	Number of Samples *p*-value < 1.0%		Number of Samples with Coefficient	Number of Samples Coefficient > 0	Number of Samples *p*-value < 1.0%
advertising	28	38	14	mobile	27	27	16
biotech	37	37	32	net hosting	8	8	6
cleantech	12	12	10	other	0	–	–
consulting	12	12	9	pub rel	10	10	6
ecommerce	12	12	6	search	0	–	–
education	0	–	–	security	0	–	–
enterprise	22	22	20	semiconductor	17	17	14
games video	28	28	16	software	90	85	55
hardware	31	31	29	web	78	70	22
legal	0	–	–				

Table A4. LDA Results of CrunchBase Data

Topic	Dimension	Top 5 Words
1	Product	video, music, digital, entertainment, artists
2	Product	news, site, blog, articles, publishing
3	Product	job, jobs, search employers, career
4	Product	people, community, members, share, friends
5	Product	facebook, friends, share, twitter, photos
6	Product	energy, power, solar, systems, water
7	Product	systems, design, applications, devices, seimconductor
8	Product	consulting, clients, support, systems experience
9	Product	event, sports, events, fans, tickets
10	Product	insurance, financial, credit, tax mortgage
11	Product	deals, shopping, consumers, local, retailers
12	Product	health, care, medical, healthcare, patient
13	Product	students, learning, education, college, school
14	Product	food, restaurants, fitnes, restaurant, pet
15	Product	investment, financial, investors, capital, trading
16	Product	advertising, publishers, advertisers, brands, digital
17	Product	manage, project, documents, document, tools
18	Product	treatment, medical, research, clinical, diseases
19	Product	games, game, gaming, virtual, entertainment
20	Product	security, compliance, secure, protection, access
21	Product	search, engine, website, seo, optimization
22	Product	search, user, engine, results, relevant
23	Product	fashion, art, brands, custom, design
24	Product	equipment, repair, car, home, accessories
25	Product	law, legal, government, public, federal
26	Product	analytics, research, analysis, intelligence, performance
27	Product	travel, travelers, vacation, hotel, hotels
28	Product	real, estate, home, buyers, property
29	Product	payment, card, cards, credit, payments
30	Technology/Product	phone, email, text, voice, messaging
31	Technology/Product	wireless, networks, communications, internet, providers
32	Technology/Product	cloud, storage, hosting, server, servers

Table A4. LDA Results of CrunchBase Data

Topic	Dimension	Top 5 Words
33	Technology/Product	app, apps, iphone, android, applications
34	Technology/Product	design, applications, application, custom, website
35	Technology/Product	site, website, free, allows, user
36	Technology/Product	testing, test, monitoring, tracking, performance
37	Market/Technology	digital, clients, brand, agency, design
38	Market	sales, customer, lead, email, leads
39	Market	solution, cost, costs, applications, enterprise
40	Market	organization, community, support, organization, businesses
41	Market	make, people, time, just, way
42	Market	quality, customer, needs, clients, provide
43	Market	systems, operates, headquartered, subsidiary, serves
44	Market	united, states, offices, america, europe
45	Market	san, york, city, california, francisco
46	Market	award, magazine, awards, bst, world
47	Market	million, world, leading, largest, global
48	Market/Team	team, experience, industry, world, market
49	Team	partners, ventures, capital, including, san
50	Team	launched, million, product, ceo, acquired