Towards A Better Measure of Business Proximity: Topic Modeling for Analyzing M&As

ZHAN SHI, Arizona State University
GENE MOO LEE, The University of Texas at Austin
ANDREW B. WHINSTON, The University of Texas at Austin

In this article, we propose a new measure of firms' dyadic business proximity. Specifically, we analyze the unstructured texts that describe firms' businesses using the natural language processing technique of topic modeling, and develop a novel business proximity measure based on the output. When compared with the existent methods, our approach provides finer granularity on quantifying firms' similarity in the spaces of product, market, and technology. We then show our measure's effectiveness through an empirical analysis using a unique dataset of recent mergers and acquisitions in the U.S. high technology industry. Building upon the literature, our model relates the likelihood of matching of two firms in a merger or acquisition transaction to their business proximity and other characteristics. We particularly employ a class of statistical network analysis methods called exponential random graph models to accommodate the relational nature of the data.

Categories and Subject Descriptors: J.4 [**Social and Behavioral Sciences**]: Economics

General Terms: Measurement, Management, Economics

Additional Key Words and Phrases: Business Proximity, Mergers and Acquisitions, Business Analytics, Topic Modeling, Exponential Random Graph Models

ACM Reference Format:
Zhan Shi, Gene Moo Lee, and Andrew B. Whinston, 2014. Towards A Better Measure of Business Proximity: Topic Modeling for Analyzing M&As *ACM* 1, 1, Article 1 (February 2014), 18 pages.
DOI: http://dx.doi.org/10.1145/0000000.0000000

1. INTRODUCTION

In this paper, we propose a text-mining-technique based measure of firms' dyadic business proximity and empirically evaluate the measure's effectiveness using a dataset of mergers and acquisitions (M&As) in the U.S. high technology (high-tech) industry. In particular, we examine the matching of companies in M&As by building statistical models that relate the likelihood of M&A between two firms to their business proximity and other characteristics.

The basic idea underlying our model is straightforward: A pair of firms that are "close" in various dimensions are more likely to be part of an M&A transaction than two that are distant. Prior research in the management, finance, and economics literature has suggested different categories of explanations why firms engage in M&A transactions: value creation, managerial self-interest (value destruction), environment factors, and firm characteristics [see Haleblian et al., 2009]. Those different antecedents have been a great inspiration for building the firm proximity measures included in our empirical model. Yet our study is not intended to argue for one particular antecedent of M&A against another, but rather, we attempt to comprehensively document the empirical evidence on the relationship between M&A likelihood and firm proximity.

Following the literature, we posit that geographic vicinity, social linkage, common ownership, and business similarity are associated with the likelihood of two high-tech firms' matching in an M&A transaction, and we construct four quantities that measure firms' dyadic proximity in these dimensions. Among the four, the most challenging has been the operationalization of business proximity, which measures firms' relatedness in the spaces of product, market, and technology. A few prior studies in

the strategic management literature have used or developed measures that serve the same or closely related purposes. Indeed, many of them adopted the same term "business proximity." The most common operationalization has been a binary variable that indicates common industry membership. With this definition, two firms' businesses are operationalized to be either identical or completely different. Stuart [1998], Mowery et al. [1998], and others constructed a "technological overlap" measure based on the 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 measure in the technology space, it requires complete data on companies' patent portfolios and does not explicitly cover the product and market spaces. Mitsuhashi and Greve [2009] focused on the market space and applied Jaccard distance on predefined geographic regions in measuring "market complementarity." A refined extension of the common industry membership definition is to use some industry classification codes in more detail. For example, in Wang and Zajac [2007], how similar two firms' businesses are was determined by the number of common consecutive digits in their industry classification codes under the North American Industrial Classification System (NAICS). Since they used the first four digits in NAICS, the similarity quantity is one of five possible values: 0.00, 0.25, 0.50, 0.75, or 1.00. The Standard Industrial Classification (SIC) codes have been similarly used by scholars in the selection of "industry rivals" [Betton et al. 2008].

In this paper, we propose a measure that can provide finer granularity in the business dimension. Using a text mining technique called topic modeling [Blei et al. 2003, Griffiths and Steyvers 2004], we analyze the unstructured texts that describe the companies' businesses. Our automatic system, the core of which is a Latent Dirichlet Allocation (LDA) algorithm, represents each company's textual description as a probabilistic distribution over a set of underlying topics, which we interpret as aspects of its businesses. Then, our business proximity can be naturally constructed by comparing a pair of firms' topic distributions. We argue that this business proximity is another step forward in measuring the closeness of companies in the arenas of product, market, and intellectual property, all of which are difficult to quantify otherwise [Baum et al. 2010].

To empirically evaluate the effectiveness of our new business proximity measure as well as to compare it with the geographic, social, and investor proximity measures in explaining M&As, we adopt a class of statistical network models called Exponential Random Graph Models (ERGMs). This modeling framework allows us to examine, among all pairs of companies, which subset of them would likely engage in M&A transactions, based on factors including but not limited to both the company-specific (nodal) characteristics and the pairwise (dyadic) relationships. The critical reason why we choose ERGMs over the conventional binary outcome econometric models such as logistic regression is that ERGMs relax the assumption of independence across different transactions. This is especially important in the M&A context where independence is clearly violated — for instance, one company cannot be acquired by two different companies.

In essence, our approach abstracts the M&As as a network — companies are nodes and transactions are edges linking the nodes, and analyzes its structure using a statistical network method. Manne [1965] viewed M&As as transactions in a "market for corporate control." In support of using the network approach to analyze markets, Jackson [2010, pg. 13] pointed out most markets "function not as centralized and anonymous institutions, but rather involve a variety of bilateral exchanges or contracts." In fact, it has already been recognized in the literature that network theories and methods can be fruitfully applied to analyzing a variety of economic exchanges and markets, for example international trade, strategic alliance, and inter-bank loans [Easley and Kleinberg 2010]. However, much more effort from this stream of management lit-

erature has been paid to studying the effects of network structure than studying the network structure itself. Thus our work contributes to this under-explored area. To our knowledge, we are the first to apply ERGMs in analyzing M&As, or networks defined by economic transactions in general.

We use a unique dataset on the U.S. high-tech industry which contains the M&A transactions over a 5-year period from 2008 to 2012. This industry is characterized by significant geographic clustering (at a handful of high-tech hubs), large number of early-stage startups, rapid job mobility, high concentration of ownership at the company level, strong influence of angel and venture investors, and comparatively large volume of M&A activities. Yet, empirical research on matching in M&As in the high-tech industry has thus far been limited. In fact, the overall vast majority of M&A research has focused on larger, public corporations [Haleblian et al. 2009]. This unbalanced research development is probably due to the lack of good quality data on small, privately-held companies and the difficulty in empirically modeling matching. Our study thus serves as one of the first attempts in the M&A literature to systematically document the empirical evidence of matching in M&As in the high-tech industry. We find that our business proximity measure is positively associated with the matching likelihood and the evidence on its statistical significance is the strongest compared with proximity measured in the other dimensions. Interestingly in our dataset, geographic proximity appears to be insignificant in identifying the high-tech firms' matching in M&As.

Our paper also contributes to the rapidly growing stream of literature that leverages data science techniques in examining huge datasets for econometric modeling and/or business analytics [Choi and Varian 2012, Einav and Levin 2013, Ghose et al. 2012]. Recent years have seen a tremendous growth in the U.S. high-tech industry. One of the defining phenomena of this expansion period is an "entrepreneurial boom" characterized by the explosion of digital startups.[1] Along with this boom, not surprisingly, the media is 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 (as an alternative to internal R&D), and talented employees.[2] Venture capitalists also arrange mergers between their partially owned startups in order to consolidate resources and reduce competitive pressure.[3] The fierce competitions in both demand and supply instantaneously create the question of matching between an acquirer and a potential target in the M&A market, as the value (or disvalue) of an M&A critically depends on the synergy of their businesses and competitive strategy. A related problem is the search for targets. While almost everyone knows who the top competitors are in an industry, finding the small companies with innovative products or technology is very difficult and time consuming. We believe data analytics can contribute to alleviating some of problems in matching and search. It is reported that many of the M&A players have already been investing heavily in their analytic capacity and capability for identifying the win-win matches by rendering the decision-making processes more "data-analytics-driven".[4] Along these lines, our work reveals the great potential of extracting economically meaningful knowledge from unstruc-

[1] See "A Cambrian Moment," *The Economist*, January 18, 2014.
[2] See "Internet Mergers and Takeovers: Platforms upon Platforms," *The Economist*, May 25, 2013.
[3] An example is the acquisition of Summize 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*.
[4] See "Google Ventures Stresses Science of Deal, Not Art of the Deal," *New York Times*, June 23, 2013.

tured public data for industry analysis. The network approach employed in the paper also sheds light on the possibility and value of building a "social network for ventures," i.e., a two-sided platform that facilitates the identification of M&A targets and makes M&A transactions less opaque.

2. DATA

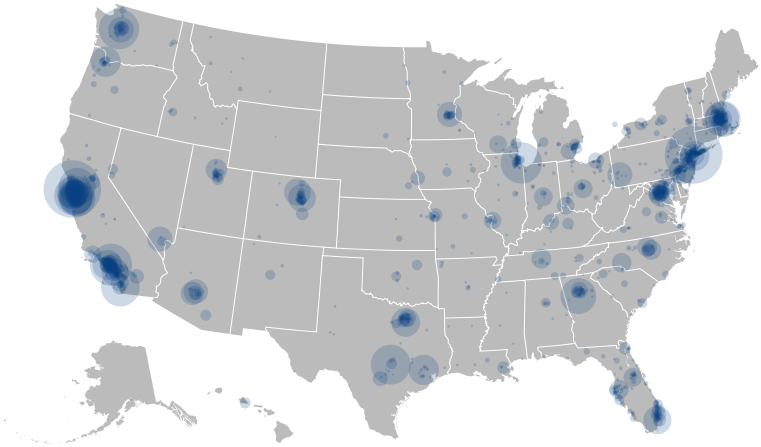
Our dataset was collected from CrunchBase[5] in April 2013. Regarded as the Wikipedia of the venture industry, CrunchBase is an open and free database of high-tech companies, people, and investors that provides a comprehensive view of the "startup world." The database automatically retrieves high-tech related information from various news sources such as `allthingsd.com`, `techcrunch.com`, and `businessinsider.com`. In addition, anyone can contribute to CrunchBase in a crowdsourcing manner. For quality assurance, each update is reviewed by moderators. Existing data is also constantly reviewed by editors.

We limit our dataset to U.S. based companies and we further exclude those for which some basic information is missing, for example a textual description. The final dataset contains $25,692$ companies. For each company, we observe its headquarter 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, and a textual description of its businesses. The unstructured textual description is mostly not very long, comprising one or more paragraphs on the key facts about the company's products, markets, and technologies. 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 headquarter 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, California leads with 34.72% of the companies, followed by New York (11.99%), Massachusetts (5.89%), Texas (5.20%), Florida (4.12%), and Washington (3.62%). We also observe an 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%).

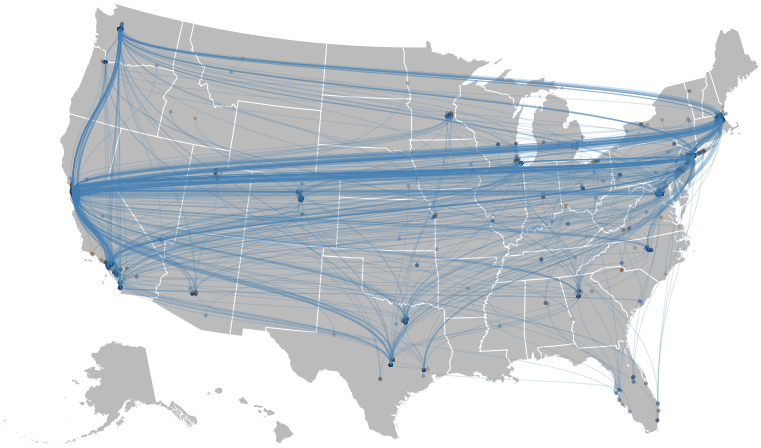
We restrict our dataset to include M&A transactions that happened in a 5-year period from 2008 to 2012. We focus on post-2008 transactions because CrunchBase was launched in late 2007 so the pre-2008 transactions were added in a retrospective manner and are more likely to be incomplete; our data collection was carried out in April 2013 so we set the end time to be the end of the previous year.[6] Overall M&As are rare events — we observe a total of $1,243$ transactions. Figure 1 (b) geo-maps each of these transactions using the headquarter locations of involved companies. Slightly less than 2/3 (62.59%) of the deals is cross-state. A numerically similar portion of transactions (63.56%) is cross-sector. The distribution of the number of transactions per company is also highly skewed — a small number of companies claim a large proportion of the transactions. 735 companies (2.86% of the total companies) have made at least one acquisition. Top 10 buyers have made 178 deals, which is 14.32% of the total M&A deals,

[5] `http://www.crunchbase.com`

[6] Hence we exclude companies that were acquired before January 1, 2008 and companies that were founded after December 31, 2012.



(a) Companies



(b) Transactions

Fig. 1: Geo-mapping Company Locations and Transactions

and top 20 contributed 21.23% of the total deals. Table V in the appendix shows the exact distribution of the number of M&A transactions per company.

3. FIRM PROXIMITY

In this section, we develop the firm proximity measures. In subsection 3.1 we describe the analytic procedure of creating a business proximity measure based on the unstructured company description data. In subsection 3.2, we discuss other firm proximity measures in the dimensions of geography, social linkage, and investment relationships.

3.1. Business Proximity

We define business proximity as a comprehensive measure of firms' closeness in the spaces of products, markets, and technologies. As discussed in the introduction, existing operationalizations used in the management, finance, and economics literature have shortcomings in classification granularity, comprehensiveness, and scalability.

Thus, our goal is to overcome limitations in these respects. Our requirement on input data is also minimal, i.e., an unstructured textual description on each firm's business. This information is much more likely to be available than structured information such as NAICS/SIC code or patent portfolio is, especially for high-tech startups.

Our approach builds upon a natural language processing technique called topic modeling. Topic modeling is a statistical model to discover abstract "topics" from a collection of documents. It is an unsupervised learning model, which means the model is automatically generated without much manual efforts in labeling each document for training. Formally, given a collection of documents, a topic model (i) discovers different topics, where each topic consists of relevant keywords, and (ii) identifies the mixture of topics in each document. The basic idea is that a specific document covers a small number of topics and the words appearing in that document are the realizations of those topics. Thus we can discover hidden topics by observing many documents. Implementations of topic modeling algorithms include Latent Semantic Analysis [Deerwester et al. 1990], Latent Dirichlet Allocation [Blei et al. 2003], and Hierarchical Dirichlet Process [Teh et al. 2006]. Among them, Latent Dirichlet Allocation (LDA) is a representative topic modeling algorithm. It has successfully applied to classify various documents including pictures, scientific articles, social network data, and survey data [see Blei 2012].

We construct our business proximity measure by applying the LDA topic modeling algorithm to the textual descriptions of firm business. Each description is a document. The algorithm produces K topics (K is specified by the researcher), where each topic is represented by a set of relevant words. In addition, LDA also outputs topic distributions for the descriptions. Specifically, for each business description, a probability value is assigned to each discovered topic and the values sum up to 1.0. Essentially, through topic modeling, each company i is represented by a topic distribution T_i.

Finally, we define the *business proximity* $p_b(i,j)$ between two companies i and j as the cosine similarity[7] of the two corresponding topic distributions T_i and T_j, which can be written as follows:

$$p_b(i,j) = \frac{T_i \cdot T_j}{||T_i||||T_j||} = \frac{\sum_{k=1}^{K} T_{i,k} T_{j,k}}{\sqrt{\sum_{k=1}^{K} (T_{i,k})^2}\sqrt{\sum_{k=1}^{K} (T_{j,k})^2}} \tag{1}$$

where $T_{i,k}$ is the k-th topic probability for company i, $k \in \{1, 2, \ldots, K\}$, and K is the total number of topics. The resulting proximity values range between 0 and 1, where a smaller value indicates closer proximity between the pair of companies.

We apply the proposed method to our dataset. We specify K to be 50. To illustrate that the topic modeling results comprehensively capture multiple dimensions of a firm's business, in Table I we list 10 topics that LDA produces from our dataset. The full 50-topic list is shown in Table VI in the appendix. We have checked all 50 topics to find that each topic consists of keywords that are tighly related to each other, while cross-topic overlaps are very small. We also observe that the topics capture the current trends in the high-tech industry.

3.2. Other Proximity Measures

3.2.1. Geographic Proximity. Geographic or spatial proximity refers to the closeness of physical locations and it has been shown to have a moderating effect in a diversity of

[7]Cosine similarity is one measure of similarity between two distributions. We can apply other similarity measures such as normalized Euclidean distance. We can also view each topic distribution as a set, and then use set comparison metrics such as Jaccard index and Dice's coefficient.

Table I: Top Words

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

financial transactions, such as mutual fund investments [Coval and Moskowitz 1999], stock tradings [Grinblatt and Keloharju 2001], bank loans [Degreyse and Ongena 2005], and venture capital financing [Sorenson and Stuart 2001]. 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 headquarter into its latitude (ϕ) and longitude (λ) coordinates 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 pairs of coordinates of two companies i and j, and $\Delta\lambda$ be the absolute difference in longitudes. Then the *geographic proximity* $p_g(i, j)$ between companies i and j is defined as

$$p_g(i, j) = -R \arccos(\sin\phi_i \sin\phi_j + \cos\phi_i \cos\phi_j \cos\Delta\lambda), \tag{2}$$

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

3.2.2. Social Proximity. Social proximity of two firms is defined based on the social linkage between the individuals associated with the two firms. Personal linkage is an important factor in coordinating transactions and promoting private information exchange between business entities through mutual trust and kinship [Hochberg et al. 2007, Cohen et al. 2008, 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 high-tech industry, especially the startup sphere of it, is characterized by job mobility, which creates the paths and opportunities for private information flow. Second, in the high-tech industry, early-stage digital startups are mostly very small in size, and thus information about them is often scarce outside the insiders' social circles. Moreover, many startups intentionally stay in a stealth mode before their products and technologies mature. To this end, we

[8] http://en.wikipedia.org/wiki/Great-circle_distance
[9] https://developers.google.com/maps/

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

$$p_s(i,j) = |S_i \cap S_j|, \tag{3}$$

i.e., the number of people who are identified having experiences in both companies.

3.2.3. Investor Proximity. Investment proximity is defined based on 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. 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 [Amit et al. 1990, Gompers 1995]. Hence, common early investors of two high-tech companies form the critical information bridge 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* $p_f(i,j)$ is defined as

$$p_f(i,j) = |I_i \cap I_j|, \tag{4}$$

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.

3.3. Analysis on Proximity Measures

In this subsection, we explore how the four proximity measures are realized in our CrunchBase dataset. Specifically, for each of the four proximity measures, we compare its different distributions in two groups of company pairs: (1) the group of M&A-matched company pairs and (2) a group of randomly-selected pairs.

Figure 2 shows the empirical cumulative distribution functions of the four proximity measures. For the (b) geographic dimension, we intentionally plot the distance rather than the proximity for intuitiveness. Also note that the business and geographic proximity values are continuous, while 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 clear distinction between the two lines, suggesting the existence of dependency between the proximity measures and M&A transactions. In the business dimension, the average proximity of M&A pairs is 0.37, 5.4 times larger than that of random pairs. In the geographic dimension, an M&A pair is on average $1,626\ km$ apart from each other, which is $518\ km$ smaller than the mean distance between a random pair. In the social dimension, a company pair linked by M&A has 0.22 common people on average, while a random pair on average has no intersection. Finally, in the investor dimension, there are 0.06 common investors between an M&A pair on average, which is 4.51 times higher than that of two randomly-paired companies.

4. EMPIRICAL ASSESSMENT

We evaluate our new business proximity measure through an empirical analysis in this section. In particular, we seek to document the relationship between the likelihood

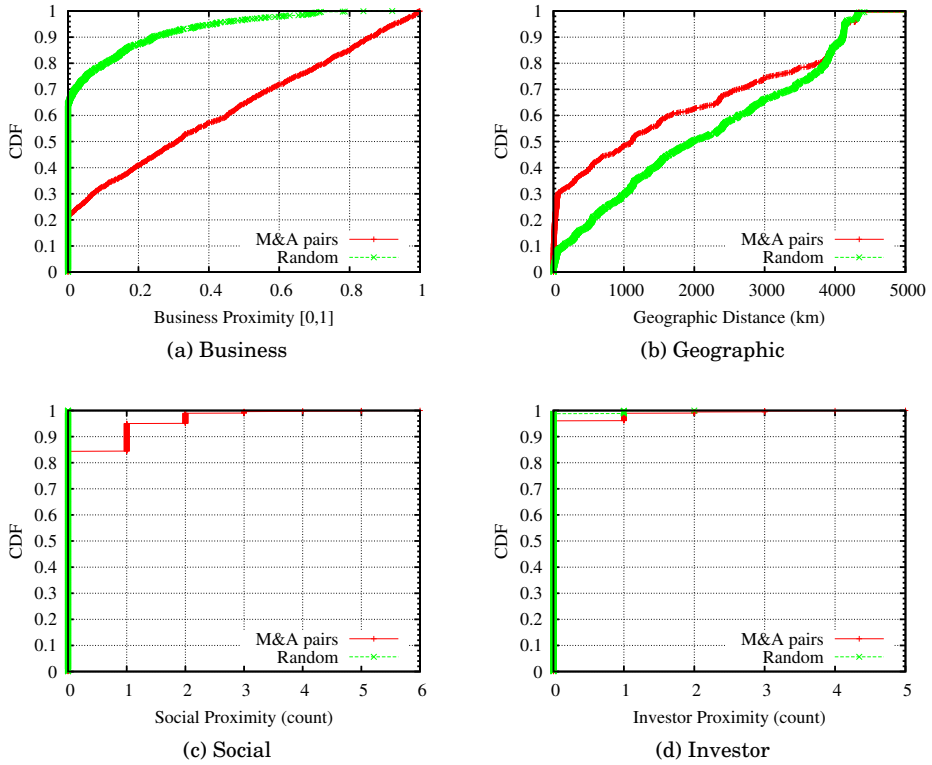


Fig. 2: Distributions of Proximity: M&A Sample v.s. Random Sample

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.

4.1. 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. However, the conventional binary response econometric models (e.g., logistic regression) are inappropriate in the present study due to the relational nature of the data. For example, an M&A transaction between firms i and j and an M&A transaction 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. 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*.

Exponential random graph models (ERGMs), a.k.a. p^* models, have been developed in statistical network analysis over the past three decades [Holland and Leinhardt 1981; Frank and Strauss 1986; Wasserman and Pattison 1996] and recently become perhaps the most important and popular class of statistical models of network structure [see Goldenberg et al. 2010]. As far as we are aware, this modeling framework has not been widely used in the management literature thus far, so we briefly introduce

it here. We also provide a list of important notations used in this and the following sections in Table VII in the appendix for easy 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 current research, the nodes and the edges are high-tech companies and the M&A transactions between them respectively, and they together form an M&A network. In terms of the adjacency-matrix representation, we define

$$Y_{ij} = \begin{cases} 1, & \text{if } i \text{ and } j \text{ are part of an M\&A transaction,} \\ 0, & \text{otherwise.} \end{cases}$$

With this definition, the resultant M&A network is undirected.[10]

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

$$\mathrm{P}(Y = y) = \frac{1}{\Psi} \exp\{\sum_{k=1}^{K} \theta_k z_k(y)\}, \tag{5}$$

where $z_k(y)$, $k = 1, 2, \ldots, K$, are K network statistics, the θ_k's are parameters, and the denominator Ψ is a normalizing constant.[11] 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 (5) in terms of log-odds of the conditional probability:

$$\mathrm{logit}(\mathrm{P}(Y_{ij} = 1 | Y_{-ij})) = \sum_{k=1}^{K} \theta_k \Delta z_k, \tag{6}$$

[10] 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 purpose 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.

[11] $\sum_{y \in \mathcal{Y}} \mathrm{P}(Y = y) = 1$, so $\Psi = \sum_{y \in \mathcal{Y}} \exp\{\sum_{k=1}^{K} \theta_k z_k(y)\}$

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.[12] [13]

4.2. 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 (6) 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 Section 2 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 has previously engaged in M&A transactions should have a different likelihood than the case where neither has. The d_2 statistic captures exactly this effect and we expect its coefficient to be positive.[14]

Selective mixing captures the matching of firms based on the combination of their *nodal-level* characteristics. In other words, these characteristics are first defined at the individual firm level, and then combined to the pair level and lastly 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," a.k.a. "homophily" in sociology. 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 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^{sec} 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

[12] 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.

[13] The above summarizes the basic formulation of ERGMs. Despite its relatively straightforward interpretation and analytic convenience, applications had been limited until just a few years ago due to significant computational burdens. The difficulty lies in evaluating the normalizing constant in the equation (5), 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 graphs is $2^{n(n-1)}$ if the network is directed, and the number of possible graphs is $2^{\frac{n(n-1)}{2}}$ if the network is undirected. Recent advances in computing capability and Monte Carlo estimation techniques [Snijders 2002, Handcock et al. 2008 among others] have made possible the significant growth of ERGMs applications in academic fields such as sociology and demography.

[14] Further, the presence of this statistic introduces dyadic dependence into our model, thereby rendering standard logistic regression inappropriate.

Table II: Degree Distribution Coefficients (100 Samples)

		Number of Samples with Coefficient	Number of Samples with Expected Sign	Number of Samples with p-value $< 1.0\%$	Median Coefficient Value
θ_t	edges	96	96(< 0)	93	-14.46
θ_{d2}	degree> 2	96	95(> 0)	67	1.67

two groups of statistics can serve as alternative operationalizations of geographic and business proximity respectively [Audretsch and Feldman 1996].

Lastly, the statistics of our most interest are the four proximity measures that capture the matching process based on *dyadic-level* characteristics. They each equal to the sum of the corresponding characteristic values over all transactions. We use p_g, p_s, p_f, and p_b to denote the sums of geographic proximity, social proximity, investor proximity, and business proximity respectively. The rationale of including them has been discussed the in Section 3.

To sum up, our model specification can be written:

$$\mathrm{P}(Y = y) = \frac{1}{\Psi} \exp\{\theta_t t + \theta_{d2} d_2 + \sum_s \theta_s^{sta} h_s^{sta} + \sum_c \theta_c^{cat} h_c^{cat} + \theta_g p_g + \theta_s p_s + \theta_f p_f + \theta_b p_b\}, \quad (7)$$

and the corresponding conditional form is

$$\begin{aligned} &\mathrm{logit}(\mathrm{P}(Y_{ij} = 1 | Y_{-ij})) \\ =&\theta_t \Delta t + \theta_{d2} \Delta d_2 + \sum_s \theta_s^{sta} \Delta h_s^{sta} + \sum_c \theta_c^{cat} \Delta h_c^{cat} + \theta_g \Delta p_g + \theta_s p_s + \theta_f \Delta p_f + \theta_b \Delta p_b \\ =&\theta_t + \theta_{d2} \Delta d_2 + \sum_s \theta_s^{sta} \mathrm{I}(s_i = s_j = s) + \sum_c \theta_c^{cat} \mathrm{I}(c_i = c_j = c) \\ &+ \theta_g p_{g,ij} + \theta_s p_{s,ij} + \theta_f p_{f,ij} + \theta_b p_{b,ij}. \end{aligned} \quad (8)$$

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

4.3. Results

The final dataset contains a total of 25,692 companies. This seemingly moderate number of nodes is actually huge for estimating network models since the number of potential edges, in our case un-ordered pairs, exceeds 330 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 for 100 times. For each of the 100 different samples (of approximately 6,400 companies each), we estimate the model coefficients by following 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 II, III, and IV respectively. For each statistic, we report out of the 100 samples the number of samples that yield a coefficient,[15] the number of samples that yield a coefficient with the expected sign, and

[15] We report no coefficient for a sample when the estimation algorithm fails to converge.

Table III: 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 Coefficient	Number of Samples Coefficient > 0	Number of Samples p-value $< 1.0\%$
AK	0	-	-	MT	4	3	2
AL	8	8	2	NC	6	6	1
AR	0	-	-	ND	0	-	-
AZ	9	9	5	NE	0	-	-
CA	100	90	17	NH	0	-	-
CO	26	26	9	NJ	45	44	15
CT	8	8	4	NM	0	-	-
DC	15	15	7	NV	0	-	-
DE	0	-	-	NY	90	72	5
FL	16	16	3	OH	16	16	5
GA	20	19	8	OK	0	-	-
HI	0	-	-	OR	0	-	-
IA	4	4	1	PA	16	15	4
ID	0	-	-	RI	0	-	-
IL	15	15	3	SC	3	3	2
IN	0	-	-	SD	0	-	-
KS	0	-	-	TN	0	-	-
KY	10	10	4	TX	64	61	11
LA	0	-	-	UT	20	20	10
MA	74	70	10	VA	32	32	10
MD	0	-	-	VT	7	7	3
ME	7	7	2	WA	57	54	10
MI	8	8	5	WI	0	-	-
MN	15	15	4	WV	0	-	-
MO	0	-	-	WY	0	-	-
MS	0	-	-				

(a) State

	Number of Samples with Coefficient	Number of Samples Coefficient > 0	Number of Samples p-value $< 1.0\%$		Number of Samples Coefficient	Number of Samples Coefficient > 0	Number of Samples p-value $< 1.0\%$
advertising	69	59	7	mobile	43	40	1
biotech	95	78	12	net hosting	54	53	19
cleantech	13	13	0	other	44	38	1
consulting	13	12	0	pub rel	16	16	0
ecommerce	66	62	11	search	5	5	2
education	0	-	-	security	37	37	15
enterprise	71	70	13	semiconductor	47	44	5
games video	75	71	12	software	100	96	47
hardware	23	23	3	web	100	89	15
legal	0	-	-				

(b) Category

Table IV: Proximity Coefficients (100 Samples)

		Number of Samples with Coefficient	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 $< 0.1\%$
θ_g	Geographic	96	49	6	1	0
θ_s	Social	96	95	69	57	45
θ_f	Investor	95	73	39	35	30
θ_b	Business	96	96	94	93	92

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 VIII.

Table II reports the coefficients of the degree distribution statistics. Among the samples that produce estimates (96 out of 100), all the θ_t coefficients (96 out of 96) are negative and all except one θ_{d2} coefficients (95 out of 96) are positive. At the 99.0% confidence level, 93 out of the 96 negative θ_t estimates are significant and 67 out of the 95 positive 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 only indicates 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 both parts (a) (b) of Table III, we observe that for almost all the 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 geographic and business proximity measures. As mentioned, the selective mixing statistics based on state and industry sector can also be thought of as alternative, but coarser operationalizations of geographic and business proximities respectively. Therefore, when including both the selective mixing statistics and our proximity measures in the ERGM specification, the effects of the selective mixing statistics are superceded by the effects of the more refined proximity measures, causing the model to produce insignificant coefficients for the selective mixing statistics. To test the validity of this explanation, we also estimate anther ERGM specification, which excludes all four proximity measures and for which we report the corresponding results for the selective mixing coefficients in Table IX in the appendix. Comparing the last columns of Table III and Table IX, we find that when using the specification without proximity measures, a much higher proportion of the samples produce statistically significant (at the 1.0% significance level) estimates for the selective mixing coefficients. This is thus a 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 IV we report the estimation results for the four proximity measures. First and foremost, the prediction that our business proximity measure is positively associated with the matching likelihood is strongly confirmed: 96 out of the 96 samples produce a positive coefficient and among them 92 estimates are significant at the 99.9% confidence level. Further, when comparing the proximity measures across the rows, we observe: The percentage of samples that yield the predicted positive coefficients ranges from 51.04% for θ_g (geographic) to 100.00% for θ_b (business); at the 95.0% confidence level, the percentages of samples that yield significantly positive coefficients are 9.38% for θ_g (geographic), 41.05% for θ_f (investor), 71.88% for θ_s (social), and 97.92% for θ_b (business); at the 99.0% confidence level, the percentages of samples that generate statistically significantly positive coefficients are 1.04% for θ_g (geographic), 36.84% for θ_f (investor), 59.38% for θ_s (social), and 96.88% for θ_b (business); at the 99.9% confidence level, the percentages of samples that generate statistically significantly positive coefficients are 0.00% for θ_g (geographic), 31.58% for θ_f (investor), 46.88% for θ_s (social), and 95.83% for θ_b (business). These results show that three among the four proximity measures (except θ_g geographic) are positively associated with the likelihood of matching in M&As. In particular, our newly developed business proximity measure also outperforms the other three measures in terms of statistical significance.

It is also noteworthy in Table IV that the geographic proximity turns out to play a less significant role in identifying high-tech firms' matching in M&As. And this result

does not seem to be caused by the simultaneous inclusion of the other three proximity measures because the weak significance of the geographic proximity is retained in an exercise, reported in Table X in the appendix, where we use each of the four proximity measures in four separate specifications (the degree distribution statistics and the selective mixing statistics are kept the same as in the main model). This is an interesting result that appears in contrast to the recent study in Chakrabarti and Mitchell [2013], who found a significant preference for geographically close targets in the acquisitions by U.S. chemical manufacturers. The different findings can probably be attributed to (1) the industry difference between high-tech and chemical (the varied costs for consolidating and integrating resources over long physical distance), and (2) the time-period difference between 1980-2003 (Chakrabarti and Mitchell 2013) and 2008-2012 (the present study). It can be an interesting future research topic to investigate how the role of geographic distance in M&As differs across industries and time.

5. DISCUSSION AND CONCLUSION

In this study we set out with the task of developing a new, more refined measure of firms' dyadic proximity in the business dimension. Through an example that uses a unique dataset of the U.S. high-tech industry, we detailed the process of topic modeling on the textual descriptions of the companies' businesses and constructing our proximity measure according to the output. We then empirically evaluated the measure's effectiveness in the context of modeling matching in M&As. In doing so, we also comprehensively documented the evidence on the relationship between the matching likelihood and high-tech firms' geographic, social, investor, and business proximities, all of which have been suggested crucial for M&As in the literature. The results demonstrated that the business proximity, as quantified by the proposed measure, is strongly associated with the matching likelihood.

We believe this research contributes to the literature in at least three very important ways with implications for both understanding and practice. First, measuring firms' relatedness in business is very important for managers to identify potential partners, competitors, and alliance or acquisition targets. However, as far as we are aware, it had not been shown that the measurement can be done in an automatic, "analytics-driven" way and at the same time provides very fine granularity. The saying in management goes, "if you cannot measure it, you cannot manage it." As shown in the paper, the new proximity measure we developed provides finer granularity in quantifying a pair of firms' relatedness in spaces such as product, market, and technology. In addition, the measure integrates the natural language processing technique of topic modeling into the operationalization of an important economic/business concept. Thus it responds to a call in the literature for incorporating machine learning techniques into the development of novel measurements (Einav and Levin 2013). More generally, this research also joins the growing stream of management literature that leverages data science in analyzing large volume of data for business analytics.

Second, the study furthers our knowledge about M&As by comprehensively documenting the empirical evidence on the relationship between the likelihood of matching and firm proximity measured in a variety of different dimensions. Moreover, our dataset on the U.S. high tech industry contains a large proportion of early-stage, private companies, which previously have not been the focus of M&A research. Thus the present study contributes to this under-explored research area. Also, the prediction that geographic proximity is important in identifying M&A targets is intriguingly not supported by our analysis, which perhaps may draw management and finance scholars to further investigate the role of geographic distance in today's business environment.

Lastly, when evaluating our business proximity measure in studying firms' matching in M&As, we adopt the statistical modeling framework of ERGMs to accommodate the

relational nature of our data. Whereas the management literature is abundant with studies on how networks affect the interaction and performance of organizations, using rigorous statistical methods to analyze the structure of inter-organizational networks is underdeveloped. To the best of our knowledge, this study is the first that applies ERGMs in the analysis of M&As, or more broadly, it is the first that uses a statistical network model to analyze relational transactions among organizations. We believe statistical network models are currently underutilized by management scholars in their empirical research on inter-organizational linkage despite the fact that relational data is actually not uncommon in the studies of many very important research questions. For example, strategic alliances, investments, and patent license agreements among companies can all be visualized and careful analyzed as graphs or networks. We predict that with the growing availability of data and the development of computing power and techniques, statistical network models' value in management research will be increasingly recognized.

Our research is not without its limitations. First, owing to the data limit, we could not empirically compare our business proximity measure with the measure based on industry classification [Wang and Zajac 2007] or the measure based on patent portfolio [Stuart 1998]. Second, some important company-level characteristics, notably company age, size, and revenue, were unavailable in our dataset, which inevitably limited our ability to extend our study. For instance, if we had 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. Third, in performing topic modeling on the companies' descriptions, we used the number of topics as a fixed parameter. While choosing one fixed number of topics is sufficient for our purpose of illustrating the process of constructing the business proximity measure, it could be practically interesting to carefully examine how the value of the constructed measure and its explanatory power vary with the choice of the number-of-topics parameter. Lastly, the model we employed in the empirical analysis can be extended or modified in a few different ways. One possibility is to use SERGMs [Chandrasekhar and Jackson 2013] to improve estimation efficiency. Secondly, the standard ERGM is a static model. To deepen our understanding about the dependence structure of M&A transactions, future research could examine the evolution of the M&A network by using some dynamic network models.

ELECTRONIC APPENDIX

The electronic appendix for this article can be accessed in the ACM Digital Library.

REFERENCES

Audretsch, D.B. and M.P. Feldman 1996, R&D Spillovers and the Geography of Innovation and Production. *American Economic Review*, 86(3), 630-640.

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

Baum, J.A.C., R. Cowan, and N. Jonard 2010, Network-Independent Partner Selection and the Evolution of Innovation Networks. *Management Science*, 56(11), 2094-2110.

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

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

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

Chakrabarti, A. and W. Mitchell 2013, The Persistent Effect of Geographic Distance in Acquisition Target Selection. *Organization Science*, 24(6), 1805-1826.
Chandrasekhar, A.G. and M.O. Jackson 2013, Tractable and Consistent Random Graph Models. `http://arxiv.org/pdf/1210.7375.pdf`.
Choi, H. and H. Varian 2012, Predicting The Present with Google Trends. *Economic Record*, 88, 2-9.
Cohen, L., A. Frazzini, and C.J. Malloy 2008, The Small World of Investing: Board Connections and Mutual Fund Returns. *Journal of Political Economy*, 116(5), 951-979.
Coval, J.D. and T.J. Moskowitz 1999, Home Bias at Home: Local Equity Preference in Domestic Portfolio. *Journal of Finance*, 54(6), 2045-2073.
Deerwester, S.C. S.T. Dumais, T.K. Landauer, G.W. Furnas, and R.A. Harshman 1990, Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41(6), 391-408.
Degreyse, H. and S. Ongena 2005, Distance, Lending Relationships, and Competition. *Journal of Finance*, 9(1), 231-266.
Easley, D. and J. Kleinberg 2010, *Networks, Crowds, and Markets: Reasoning About a Highly Connected World*. Cambridge University Press, 2010.
Einav, L. and J.D. Levin 2013, The Data Evolution and Economic Analysis. *NBER Working Paper 19035*, May 2013.
Erel, I., R.C. Liao, and M.S. Weisbach 2012, Determinants of Cross-Border Mergers and Acquisitions. *Journal of Finance*, 67(3), 1045-1082.
Frank, O. and D. Strauss 1986, Markov Graphs. *Journal of the American Statistical Association*, 81, 832-842.
Ghose, A., P.G. Ipeirotis, and B. Li 2012, Designing Ranking Systems for Hotels on Travel Search Engines by Mining User-Generated and Crowd-Sourced Content. *Marketing Science*, 31(3), 493-520.
Goldenberg, A., A.X. Zheng, S.E. Fienberg, and E.M. Airoldi 2010, A Survey of Statistical Network Models. *Foundations and Trends in Machine Learning*, 2(2), 129-233.
Gompers, P.A. 1995, Optimal Investment, Monitoring, and the Staging of Venture Capital. *Journal of Finance*, 50(5), 1461-1489.
Grinblatt, M. and M. Keloharju 2001, How Distance, Language, and Culture Influence Stockholdings and Trades. *Journal of Finance*, 56(3), 1053-1073.
Griffiths, T.L. and M. Steyvers 2004, Finding Scientific Topics. *Proceedings of the National Academy of Science*, 101, 5228-5235.
Haleblian, J., C.E. Devers, G. McNamara, M.A. Carpenter, and R.B. Davison 2009, Taking Stock of What We Know About Mergers and Acquisitions: A Review and Research Agenda. *Journal of Management*, 35(3), 469-502.
Handcock, M.S., D.R. Hunter, C.T. Butts, S.M. Goodreau, and M. Morris 2008, **statnet**: Software Tools for the Representation, Visualization, Analysis and Simulation of Network Data. *Journal of Statistical Software*, 24, 1-11.
Hochberg, Y., A. Ljungqvist, and Y. Lu 2007, Whom You Know Matters: Venture Capital Networks and Investment Performance. *Journal of Finance*, 62(1), 251-301.
Holland, P.W. and S. Leinhardt 1981, An Exponential Family of Probability Distributions for Directed Graphs. *Journal of the American Statistical Association*, 76(373), 33-50.
Jackson, M.O. 2010, *Social and Economic Networks*. Princeton University Press, 2010.
Lorenzoni, G. and A. Lipparini 1999, The Leveraging of Interfirm Relationships as A Distinctive Organizational Capability: A Longitudinal Study. *Strategic Management Journal*, 20(4), 317-338.
Manne, H.G. 1965, Mergers and the Market for Corporate Control. *Journal of Political Economy*, 73(2), 110-120.

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

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

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

Sorenson, O. and T.E. Stuart 2001, Syndication Networks and The Spatial Distribution of Venture Capital Financing. *American Journal of Sociology*, 106(6), 1546-1588.

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), 668-698.

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

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

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

Wasserman, S. and P. Pattison 1996, Logit Models and Logistic Regressions for Social Networks: I. An Introduction to Markov Graphs and p^*. *Psychometrika*, 60, 401-425.

Online Appendix to: Towards A Better Measure of Business Proximity: Topic Modeling for Analyzing M&As

ZHAN SHI, Arizona State University
GENE MOO LEE, The University of Texas at Austin
ANDREW B. WHINSTON, The University of Texas at Austin

A. ADDITIONAL TABLES

Table V: The Distribution of Number of Transactions per Company

Number of Deals	Number of Companies
0	23,775
1	1,686
2	147
3	33
4	16
5	11
6	4
7	2
8	2
9	4
10	2
11	1
14	4
15	1
18	1
21	1
24	1
33	1

DOI: http://dx.doi.org/10.1145/0000000.0000000

Table VI: Top Words

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,semiconductor
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,fitness,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
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	organizations,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,best,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,acquirede

Table VII: Notations

Network graph	
Y, Y_{ij}	a random network graph matrix, its i, j element
Y_{-ij}	all elements except i, j
$\mathcal{Y}$	the set of all possible graphs for a fixed set of nodes
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 headquarter 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 VIII: Model Coefficients from Sample 1

	Coeff	S.E.	p-value		Coeff	S.E.	p-value
Geographic	0.0409	0.0272	0.1323	NV	-	-	-
Social	2.0551	0.9138	0.0245	NY	0.7842	0.8714	0.3681
Investor	0.1229	0.1809	0.4971	OH	3.8046	2.3563	0.1064
Business	0.0465	0.0046	0.0000	OK	-	-	-
Edges	-17.6608	2.4243	0.0000	OR	-	-	-
Degree>2	1.8238	0.4169	0.0000	PA	-	-	-
State				RI	-	-	-
AL	-	-	-	SC	-	-	-
AR	-	-	-	SD	-	-	-
AZ	-	-	-	TN	-	-	-
CA	0.5776	0.4289	0.1780	TX	1.5709	1.4750	0.2869
CO	-	-	-	UT	-	-	-
CT	-	-	-	VA	-	-	-
DC	5.7309	7.3488	0.4355	VT	-	-	-
DE	-	-	-	WA	0.8628	2.7314	0.7521
FL	-	-	-	WI	-	-	-
GA	-	-	-	WV	-	-	-
HI	-	-	-	WY	-	-	-
IA	-	-	-	Category			
ID	-	-	-	advertising	0.7676	1.3611	0.5728
IL	-	-	-	biotech	1.2036	1.0375	0.2460
IN	-	-	-	cleantech	-	-	-
KS	-	-	-	consulting	1.2023	1.7029	0.4802
KY	-	-	-	ecommerce	2.0914	0.9799	0.0328
LA	-	-	-	education	-	-	-
MA	-	-	-	enterprise	-	-	-
MD	-	-	-	games video	0.8704	1.5792	0.5815
ME	-	-	-	hardware	-	-	-
MI	-	-	-	legal	-	-	-
MN	-	-	-	mobile	-	-	-
MO	-	-	-	network hosting	-	-	-
MS	-	-	-	other	0.7519	1.0248	0.4631
MT	-	-	-	public relations	-	-	-
NC	-	-	-	search	-	-	-
NE	-	-	-	security	-	-	-
NH	-	-	-	semiconductor	2.6170	2.3680	0.2691
NJ	-	-	-	software	1.4763	0.4501	0.0010
NM	-	-	-	web	0.8147	0.6123	0.1834

Table IX: Selective Mixing Coefficients (100 Samples): One Specification Excluding All Proximities

	Number of Samples with Coefficients	Number of Samples Coefficient > 0	Number of Samples p-value $< 1.0\%$		Number of Samples Coefficient	Number of Samples Coefficient > 0	Number of Samples p-value $< 1.0\%$
AK	0	-	-	MT	4	4	2
AL	8	8	7	NC	6	6	4
AR	0	-	-	ND	0	-	-
AZ	9	9	7	NE	0	-	-
CA	100	100	81	NH	0	-	-
CO	26	26	25	NJ	45	45	39
CT	8	8	8	NM	0	-	-
DC	15	15	15	NV	0	-	-
DE	0	-	-	NY	90	89	22
FL	16	16	3	OH	16	16	16
GA	20	20	18	OK	0	-	-
HI	0	-	-	OR	0	-	-
IA	4	3	1	PA	16	16	16
ID	0	-	-	RI	0	-	-
IL	15	15	11	SC	3	3	2
IN	0	-	-	SD	0	-	-
KS	0	-	-	TN	0	-	-
KY	10	10	10	TX	64	64	23
LA	0	-	-	UT	20	20	20
MA	74	74	32	VA	32	32	32
MD	0	-	-	VT	7	7	2
ME	7	7	4	WA	57	57	35
MI	8	8	8	WI	0	-	-
MN	15	15	13	WV	0	-	-
MO	0	-	-	WY	0	-	-
MS	0	-	-				

(a) State

	Number of Samples with Coefficient	Number of Samples Coefficient > 0	Number of Samples p-value $< 1.0\%$		Number of Samples Coefficient	Number of Samples Coefficient > 0	Number of Samples p-value $< 1.0\%$
advertising	69	69	34	mobile	43	43	11
biotech	95	95	82	net hosting	54	54	54
cleantech	13	13	13	other	44	44	3
consulting	13	12	0	pub rel	16	16	15
ecommerce	66	66	31	search	5	5	5
education	0	-	-	security	37	37	37
enterprise	71	71	41	semiconductor	47	47	47
games video	75	75	42	software	100	100	84
hardware	23	23	20	web	100	96	49
legal	0	-	-				

(b) Category

Table X: Proximity Coefficients (100 Samples): Four Specifications Each with One Proximity

		Number of Samples with Coefficient	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 $< 0.1\%$
θ_g	Geographic	100	61	13	5	1
θ_s	Social	100	91	87	78	71
θ_f	Investor	100	95	73	65	49
θ_b	Business	100	100	100	100	100