Supporting Information

Atalay et al. 10.1073/pnas.1015564108

SI Text

SI Text contains five sections. In the first section, we discuss Eqs. **1–4**. In the second section, we briefly describe the dataset. In the third section, we provide a sensitivity analysis of our maximum likelihood estimate of r. In the fourth section, we present the results of a series of regressions, which characterize the probability that two firms are linked to one another. Finally, in the fifth section, we examine the importance of one potential source of sample selection bias.

Eqs. 1–4

Eqs. 1 and 2. Eq. **1** in the main text is (Eq. **S1**)

$$\frac{\partial}{\partial t}n(k,t)+\frac{\partial}{\partial k}[n(k,t)\gamma(k)] = \beta(k,t)N(t)(q+g)-qn(k,t). \quad [S1]$$

In this equation, $n(k,t)$ is the number of vertices with in-degree k at time t, $\gamma(k)\equiv\frac{dk}{dt}$ is the rate at which vertices of in-degree k gain new predecessors, q is the rate at which vertices leave the network, g is the growth rate of the number of vertices in the network, and $\beta(k,t)$ is the in-degree distribution for entering vertices. Finally, $N(t)=\int n(k,t)dk$ is the total number of vertices in the network at time t.

Eq. **S1** is analogous to equation 3 in ref. 1. This partial differential equation (PDE) describes the evolution of the distribution of in-degrees across time. Eq. **S1** is a special case of the Forward Kolmogorov Equation in which there is no variability in the growth of the in-degree of an existing vertex. For example, see ref. 2 or section 3.4 in ref. 3 for a discussion of the Forward Kolmogorov Equation. To derive Eq. **S1**, one could follow an argument given on pages 915–917 of ref. 4. This argument involves counting the number of vertices with an in-degree between k_0 and k_1 at times t and $t+\Delta$ for some small, positive Δ. As Δ approaches 0, the relationship between the number of vertices at time t and $t+\Delta$ approaches Eq. **1** of the main text.

To arrive at Eq. **2** of the main text, use the definition of $p(k,t)\equiv\frac{n(k,t)}{N(t)}$ and the product rule (Eq. **S2**):

$$\frac{\partial p(k,t)}{\partial t}N(t)+\frac{dN(t)}{dt}p(k,t)+\frac{\partial(\gamma(k)n(k,t))}{\partial k} = \beta(k,t)N(t)(q+g)-qn(k,t). \quad [S2]$$

Dividing by the total number of nodes at time t and rearranging produces (Eq. **S3**)

$$\frac{\partial p(k,t)}{\partial t}+\frac{\partial(\gamma(k)p(k,t))}{\partial k} = \beta(k,t)(q+g)-\left(q+\frac{\dot{N}(t)}{N(t)}\right)p(k,t). \quad [S3]$$

Because $\frac{\dot{N}(t)}{N(t)}=g$, Eq. **S3** is equivalent to Eq. **2** in the main text.

Solving Eq. 3. Eq. **3** in the main text is (Eq. **S4**)

$$\frac{\partial}{\partial k}\left[p(k)\left(qr(m-k)+\frac{\delta(k+r(m-k))(q+g)}{1-q}\right)\right] = (q+g)\left(\frac{e^{-\frac{k}{m(1-\delta)}}}{m(1-\delta)}-p(k)\right). \quad [S4]$$

To solve Eq. **S4**, we first rearrange terms (Eqs. **S5** and **S6**):

$$p'(k)\left(qr(m-k)+\frac{\delta(k+r(m-k))(q+g)}{1-q}\right)+p(k)\left[-qr+\frac{\delta(q+g)}{1-q}(1-r)+(q+g)\right] = (q+g)\frac{e^{-\frac{k}{m(1-\delta)}}}{m(1-\delta)} \quad [S5]$$

and

$$p'(k)+p(k)\frac{-qr(1-q)+\delta(q+g)(1-r)+(q+g)(1-q)}{qr(m-k)(1-q)+\delta(k+r(m-k))(q+g)} = \frac{\exp\left\{-\frac{k}{m(1-\delta)}\right\}(q+g)(1-q)}{m(1-\delta)qr(m-k)(1-q)+(q+g)m(1-\delta)\delta(k+r(m-k))}. \quad [S6]$$

The in-degree distribution is described by a linear first-order differential equation. A linear first-order differential equation of the form $y'(x)+f_1(x)y(x)=f_0(x)$ has a solution given by $y(x)=e^{-\int f_1(x)dx}\left(\int f_0(x)e^{\int f_1(x)dx}dx+\kappa\right)$; κ is a constant of integration.

Simple calculations yield (Eqs. **S7** and **S8**)

$$\exp\left\{-\int\frac{-qr+\frac{\delta(q+g)}{1-q}(1-r)+(q+g)}{qr(m-k)+\frac{\delta(k+r(m-k))(q+g)}{1-q}}dk\right\} = \lambda_0(k+R)^{-1-S} \quad [S7]$$

and

$$\int\left\{\exp\left\{\int\frac{-qr+\frac{\delta(q+g)}{1-q}(1-r)+(q+g)}{qr(m-k)+\frac{\delta(k+r(m-k))(q+g)}{1-q}}dk\right\}\times\frac{\frac{e^{-\frac{k}{m(1-\delta)}}}{m(1-\delta)}}{\frac{qr(m-k)}{q+g}+\frac{\delta(k+r(m-k))}{1-q}}\right\}dk = -\lambda_1\Gamma\left[1+S,\frac{R+k}{m(1-\delta)}\right]. \quad [S8]$$

As in the main text, $R=m\frac{\delta(q+g)r+qr(1-q)}{\delta(q+g)(1-r)-qr(1-q)}$, $S=\frac{(q+g)(1-q)}{\delta(1-r)(q+g)-qr(1-q)}$, and Γ is the upper incomplete γ-function; λ_0 and λ_1 are constants that do not depend on k.

Multiplying the last two terms gives us (Eq. **S9**)

$$p(k) = \lambda_0\lambda_1(k+R)^{-1-S}\left(\frac{\kappa}{\lambda_1} - \Gamma\left[1+S, \frac{R+k}{m(1-\delta)}\right]\right). \quad \textbf{[S9]}$$

The constant of integration, κ, is chosen so that the term in parentheses equals 0 for some $\hat{k}$ arbitrarily close to 0. This will ensure that $p(\hat{k}) = 0$ for the $\hat{k}$ arbitrarily close to 0.

Given this constant of integration, Eq. **S9** yields (Eq. **S10**)*

$$p(k) \propto (k+R)^{-1-S}\left(\Gamma\left[1+S, \frac{R}{m(1-\delta)}\right] - \Gamma\left[1+S, \frac{R+k}{m(1-\delta)}\right]\right). \quad \textbf{[S10]}$$

This is Eq. **4** in the main text.

Data

The dataset consists of variables drawn from the Center for Research in Security Prices (CRSP)/Compustat database.

Statement of Financial Accounting Standards (SFAS) regulation number 131, passed by the Financial Accounting Standards Board in 1997, requires publicly traded firms to report sales to customers that make up greater than 10% of the firm's total revenues in a given calendar year.† Firms are allowed to, and sometimes do, report customers that make up less than 10% of the firm's revenues. For the most part, the 10% requirement means that we observe only one or two customers for a given firm. Different firms report their customers in different ways (for example, a firm reporting General Motors as an important customer may write GM, General Motors, or Gen Mtrs). To construct our network of supplier–buyer relationships, we must use a name-matching algorithm that assigns each reported customer to a unique identifying number. This algorithm produces 39,815 firm-year observations, with 14,204 unique buyer–supplier relationships. Additional discussion of this dataset is in section 2 of ref. 5.‡

In Fig. S1, we plot the number of firms in our sample as well as the average number of suppliers per firm. The number of firms in our sample increased steadily for most of the sample period, from 631 in 1979 to 1,848 in 2002. Although the number of firms in our sample increased over time, the average number of suppliers per firm has remained fairly constant. The average number of suppliers per firm was slightly above one throughout the sample period. Most firms in our dataset, however, are not reported as customers by other firms. Conditional on being reported as a customer by at least one firm, the number of suppliers per firm is 3.67 during our sample period.

In addition to the data on firms' customers, we include information on the location of the firms' headquarters, the number of employees, total sales, and the firm's four-digit standard industrial classification (SIC) industry. In Table S1, we list the industries in our dataset that account for the largest share of revenues.

Manufacturing firms are overrepresented in our dataset. The total revenue for firms in our dataset was $4.47 trillion in 1997; aggregate gross output for the United States was $14.86 trillion in that year. Slightly greater than one-half of the $4.47 trillion can be attributed to firms in the manufacturing sector. For the United States, only one-quarter ($3.73 trillion) of gross output was earned by firms in the manufacturing sector.

Sensitivity Analysis of Maximum Likelihood Estimation

Of the model's five parameters, only r cannot be estimated by computing a sample mean of the microdata. The parameter is the fraction of new edges that are assigned randomly among the existing vertices. For a given value of r, the probability that a vertex with in-degree k receives a given edge is $r\frac{1}{N} + (1-r)\frac{k}{mN}$. With probability r, the new edge is assigned with equal probability to one of the N incumbent vertices. With probability $1-r$, the new edge is assigned by a preferential attachment rule. Under this preferential attachment rule, the probability that a vertex with in-degree k receives the new edge is $\frac{k}{m \cdot N}$.

The maximum likelihood estimate of r is the value, restricted to be in the unit interval, that maximizes (Eq. **S11**)

$$\mathcal{L}(r) = \sum_{\text{new edges}} \log\left(r\frac{1}{N(t-1)} + (1-r)\frac{k_i(t-1)}{m(t-1)\cdot N(t-1)}\right). \quad \textbf{[S11]}$$

In Eq. **S11**, $k_i(t-1)$ is the in-degree (at time $t-1$) of the vertex that receives the edge, $m(t-1)$ is the average in-degree for vertices that survive from period $t-1$ to t, and $N(t-1)$ is the number of vertices that survive from period $t-1$ to period t. When computing $k_i(t-1)$, instead of counting the total number of edges that i is receiving at time $t-1$, we only count the edges that are present in $t-1$ and are also present in period t. This accords with the timing of our network formation model: first, vertices lose some of their suppliers, and then, new edges form, some randomly and some under preferential attachment. Because the new edges form after suppliers are lost, we should not count the lost partners when computing the in-degree for a vertex (which determines the probability of forming new edges). As we reported in the main text, the maximum likelihood estimate of r is 0.18. This figure can be read off Table S2 in the last column of the first row.

In the other cells of Table S2, we estimate r from Eq. **S11** for different subsamples of the dataset. In our model, the fraction of $j \to i$ edges that are randomly assigned is the same both for edges with j as an entering firm and edges with j as an incumbent firm. In the first column of Table S2, we estimate r using only data for new edges from existing vertices to existing vertices. In the second column, we estimate r using data for new edges, where the originating vertex was not present in the previous year. The estimated randomness coefficient is somewhat higher for new vertex to existing vertex edges. We also estimate r for different

*The constant of proportionality in the solution to the PDE is $\exp\{\frac{R}{m(1-\delta)}\}S(m(1-\delta))^S$. With this constant of proportionality in hand, we can check that $\int_0^\infty p(k)dk = 1$ and that Eqs. **S5** and **S6** hold. First, $\int_0^\infty p(k)dk = 1$ holds, because

$$\int_0^\infty (k+R)^{-1-S}\Gamma\left[1+S, \frac{R}{m(1-\delta)}\right]dk = \Gamma\left[1+S, \frac{R}{m(1-\delta)}\right]\frac{R^{-S}}{S}$$

and

$$\int_0^\infty (k+R)^{-1-S}\Gamma\left[1+S, \frac{R}{m(1-\delta)}\right]dk = \frac{\exp\left\{-\frac{R}{m(1-\delta)}\right\}(m(1-\delta))^{-S}}{S} - \Gamma\left[1+S, \frac{R}{m(1-\delta)}\right]\frac{R^{-S}}{S}.$$

Second, combine the following terms,

$$p'(k) = (m(1-\delta))^S(k+R)^{-2-S}S\frac{\exp\left\{-\frac{k}{m(1-\delta)}\right\}(k+R)^{S+1}}{((1-\delta)m)^{S+1}} - (m(1-\delta))^S(k+R)^{-2-S}S\exp\left\{\frac{R}{m(1-\delta)}\right\}(1+S) \times\left(\Gamma\left[1+S, \frac{R}{m(1-\delta)}\right] - \Gamma\left[1+S, \frac{R+k}{m(1-\delta)}\right]\right)$$

and

$$p(k)\frac{-qr + \frac{\delta(q+g)}{1-q}(1-r) + (q+g)}{qr(m-k) + \frac{\delta(k+r(m-k))(q+g)}{1-q}} = \frac{\left(\exp\left\{\frac{R}{m(1-\delta)}\right\}\left(\Gamma\left[1+S, \frac{R}{m(1-\delta)}\right] - \Gamma\left[1+S, \frac{R+k}{m(1-\delta)}\right]\right)\right)}{\delta k(q+g) + (m-k)(q(1-q) + \delta(q+g))r} \times (k+R)^{-1-S}S\exp\left\{\frac{R}{m(1-\delta)}\right\}(m(1-\delta))^S \times [g(1-q+\delta(1-r)) + q(1-r)(1-q+\delta)]$$

to see that Eqs. **S5** and **S6** hold.

†SFAS 14, which was passed in 1977, also required publicly traded firms to report sales to any customers that make up more than 10% of revenues.

‡With data from 1980 to 2004, Cohen and Frazzini (5), using a similar algorithm, are able to create a dataset with 11,484 unique buyer–supplier relationships.

time periods. The randomness coefficients are only slightly lower in the first part of our sample period.

Probability of Link Formation

In this section, we describe the results of a series of logit regressions. The aim of these regressions is to find variables that are useful in predicting whether two vertices are linked to one another. Our model predicts that the in-degree of vertex j in period $t-1$ is positively related to the probability that an edge forms between vertex i and vertex j in year t. Of course, there are other variables that could potentially determine whether two firms are likely to interact with each other.

One set of variables that we use measures how similar firms are to one another. One such measure of similarity is the physical distance between firms. Ref. 6 has a review of the literature on gravity equations—equations that are used to estimate the effect of distance on the amount of aggregate trade between countries. Disdier and Head (7) perform a metaanalysis of over 100 separate papers that estimate a gravity equation to determine how the estimates of the effect of distance on trade flows have changed over time. A complementary set of papers uses microdata to study the extent to which distance reduces the probability that two individuals or firms will interact. For instance, the probability that a given individual wins an eBay auction is significantly higher when the buyer and seller are in the same city (8). Distance, interpreted loosely, can measure individuals' dissimilarity along dimensions other than physical location. Using a dataset on high school friendships, Currarini et al. (9) document that students are significantly more likely to form connections with their peers from the same race. In our logit regressions, we will include not only the physical distance between firms as an explanatory variable but also a set of indicator variables describing whether the two firms are in the same industry.

It is also possible that two individuals are more likely to be connected when they share a common contact. This is the case in several social networks (10). We include, as an explanatory variable, the number of vertices, k, such that i sells to k and k sells to j in year t.

In Table S3, we present the results of our logit regressions. The dependent variable is the probability that firm i sells to firm j in a given year, $t \in \{1997, \ldots 2007\}$. The regression sample includes all ij pairs, where i is a seller and j is a buyer in the given year. In this sample, an edge exists for 0.26% of the ij pairs. We find that the number of customers of firm j in year $t-1$ is indeed an important predictor of the probability that an ij edge exists. According to the model given in the penultimate column of Table S3, 1 SD of the previous in-degree of the customer is associated with a 0.03% higher probability that an edge exists. An additional common partner, k, is associated with a 0.12% increase in the probability that an edge exists between vertices i and j.

Distance is also an important determinant of the probability that two vertices are linked to one another. We measure the physical distance between two firms as the great circle distance between the headquarters of the two firms. Compared with firm-pairs for which the supplier and customer are 100–500 mi apart, two firms with headquarters less than 25 mi apart are 0.18% more likely to be connected.

Firms that are in the same industry are more likely to interact with one another. Compared with firms that are not in the same one-digit SIC industry, firms that are in the same two-digit industry have a 0.36% higher probability of buying from one another; firms in the same three-digit industry have a 1.22% higher probability of buying from one another, and firms in the same four-digit industry have a 1.69% higher probability of buying from each other.

Table S4 presents the results from a related series of logit regressions. Instead of running the regressions on a sample of all ij pairs such that i is a seller and j is a buyer, we only consider pairs where i is an entering firm (i.e., not present in the network in the previous year). The estimated marginal effects are, in general, similar to those reported in Table S3. A 1-SD increase in the year $t-1$ in-degree of vertex j is associated with a 0.03% higher probability that firm i sells to firm j. Compared with firms that are located 100–500 mi apart, firms that have headquarters that are less than 25 mi apart are 0.18% more likely to be linked. Compared with two firms that are not in the same industry, firms in the same four-digit SIC are 1.87% more likely to be linked to one another.

How Important Is the 10% Cut-Off Rule?

Introduction. Because of the way firms report who their customers are, one may be concerned that we are undercounting the number of suppliers, especially for small firms. Firms are told to report all firms that account for at least 10% of their sales in a given year. In Fig. S2, we see that there are some firms that do report customers that account for less than 10% of sales. However, the 10% rule does have a large effect on the number of observed edges.

To determine whether the undercounting problem is more severe for small firms, we will use the following two pieces of information:

i) The right tail of the distribution of link value as a fraction of supplier's sales. We will try to extrapolate—using information about the distribution to the right of the 10% cut-off—how many missing edges there are to the left of the 10% cut-off.

ii) The characteristics of the firms in the edges with values to the left and the right of the cut-off. Suppose, for example, that we find that observed edges where the customer is of below average size are not more likely to fall below the 10% cut-off. If this were the case, we would be justified in arguing that the number of unreported edges is not greater for small firms. To the extent that observed edges are more likely to be small when the customer is small, we will have evidence of more missing edges for small vs. large firms.

In the remaining parts of this section, we will make the intuition of the last two bullet points more precise. We proceed in three steps. First, we will argue that the probability that an existing edge is observed is only a function of the size of the edge as a fraction of the supplier's sales. Second, through extrapolation, we form an estimate of how many total missing edges exist. Third, we use an estimate of the effect of the size of the customer on the size of the edge to calculate how severely we are undercounting in-degree for small and large firms. To preview the main result, we find that existing edges where the customer is 1 SD above the average size (measured by log employees) are likely to be observed roughly 64% of the time. Edges where the customer is 1 SD below the average size are likely to be observed roughly 50% of the time. In other words, the 10% cut-off is causing us to undercount in-degree ~30% (~64/50 − 1) more for small firms relative to big firms. Because the rate of undercounting is similar for small and large firms, the shape of the in-degree distribution changes little when we account for the unobserved edges (Fig. S4).

Probability That an Existing Edge Is Observed Is Only a Function of the Size of the Edge as a Fraction of the Supplier's Sales. In this subsection, we argue that the probability that an edge between two firms is observed in our dataset depends only on s_{ij}, the value of the ij edge as a fraction of the supplier's total sales. Throughout, we will use the following notation. Let $i \to j$ denote the event that firm i supplies firm j. Let "observed $i \to j$" denote the event that firm i supplies firm j and that this edge is observed in the dataset. Finally, let X_{ij} denote observable characteristics of the ij pair.

Define $\psi(s,X) \equiv \Pr[\textit{observed } i \to j \mid i \to j, s_{ij} = s, X_{ij} = X]$. We want to argue that $\psi(s, X_{ij})$ does not depend on X_{ij}. In other

words, given that we know s_{ij}, no other characteristics of the supplier–buyer pairs affect whether this edge is observed. For example, conditioning on s_{ij}, the following variables will have no effect on whether the edge is observed:

i) the industry of the supplier or customer,
ii) the size of the supplier or customer, and
iii) the physical distance between the supplier and the customer.

We assume that $\psi(s, X) = 1$ for $s > 0.1$: all firms in the dataset follow regulations and list all of their customers that make up at least 10% of their sales. For $s \in (0, 0.1)$, we assume that $\psi(s, X)$ is additively separable in s and X: $\psi(s, X) = \phi(s) + \zeta(X)$.

Using Bayes' Rule, the probability that an edge is reported, given that it exists, is equal to (Eq. **S12**)

$$\psi(s,X) = \frac{\Pr[s_{ij} = s, \text{and is reported} \mid X_{ij} = X]}{\Pr[s_{ij} = s \mid X_{ij} = X]}. \qquad \textbf{[S12]}$$

Consider edges with share $s \in [0.1 - \varepsilon, 0.1 + \varepsilon]$, with ε small. On the left half of the interval, $\Pr[s_{ij} = s]$ and is reported $\mid X_{ij} = X \sim \psi(0.1, X_{ij}) \Pr[s_{ij} = s \mid X_{ij} = X]$.

On the right half of the interval, $\Pr[s_{ij} = s]$ and is reported $\mid X_{ij} = X] = \Pr[s_{ij} = s \mid X_{ij} = X]$.

If ε is small enough, throughout the interval, $s \in [0.1 - \varepsilon, 0.1 + \varepsilon]$, $\Pr[s_{ij} = s \mid X_{ij} = X]$ is just some function of X_{ij} and independent of s.

Thus, $\frac{\Pr[s_{ij} \in [0.1-\varepsilon, 0.1], \text{and is reported} \mid X_{ij}=X]}{\Pr[s_{ij} \in [0.1, 0.1+\varepsilon], \text{and is reported} \mid X_{ij}=X]} \approx \psi(0.1, X_{ij})$. However, the left-hand side is estimable using data only on reported edges. In particular, we take observed edges with s_{ij} close to 0.1 and examine whether edges are more likely to be on one side of the $[0.1 - \varepsilon, 0.1 + \varepsilon]$ interval as firm-pair characteristics change.

For $s_{ij} \in [0.09, 0.11]$, we run a logit regression using a sample of observed edges between 1979 and 2007. The dependent variable is equal to 1 if $s_{ij} \geq 0.1$ and 0 otherwise. The independent variables that we include are:

i) log employment of the supplier and customer,
ii) log (real) assets of the supplier and customer,
iii) industry (according to one-digit SIC code) of the supplier and customer,
iv) the physical distance between the supplier and customer,
v) whether the two firms share the same one-digit industry,
vi) a trend variable.

In Table S5, we present the coefficient estimates and robust SEs. Except perhaps for distance, none of the explanatory variables are statistically significant: $\psi(0.1, X)$ does not depend on any of the covariates that we chose. Therefore, $\psi(0.1, X) = \phi(0.1) + C$ for some constant C and some function ϕ. Because of the additive separability that we assumed for $\psi(s, X)$, $\psi(s, X) = \phi(s) + C$.

To summarize this subsection, we defined a function ψ. This function gives the probability of an edge with $s_{ij} = s$ observed, conditional on (*i*) the edge of size s existing and (*ii*) other characteristics of the ij firms. We assumed that ψ was additively separable in s and X. Given this assumption, we showed that, conditional on s_{ij}, buyer–supplier characteristics are not important in explaining whether an edge is observed, given that it exists.

How Many Missing Edges Are There to the Left of the 10% Cut-Off? One problem is that we do not know what $\phi(s)$ looks like. There should be some way to make inferences about ϕ as we observe all edges for $s \geq 0.1$ and some edges for $s \in (0, 0.1)$. Suppose that edge values, s, are distributed according to a random variable that has an associated probability distribution function f. We observe f only for $s \in [0.1, 1]$. Our estimates of $f_m(s)$ for $s < 0.1$ are the predicted values of an Epanechnikov kernel-weighted local mth-order polynomial regression using data from $s \geq 0.1$. As we can see in Fig. S3, involving higher-order terms in the polynomial regression increases our estimate of f over the (0, 0.1) interval.

Table S6 gives $f_m(s)$ and $\phi_m(s) \equiv \frac{f_m(s)}{\text{Number of edges with value in bin s}}$ for $m = 0, 1, 2$. This table is simply a second way to visualize the data in Fig. S3.

Therefore, for example, based on our extrapolation, we estimate that roughly 10–25% of the edges that have $s_{ij} \in (1\%, 2\%)$ are reported. We will use ϕ_2 later on, because this provides the most conservative estimate of the number of reported edges.

Is an Edge from Firm *i* to Firm *j* Less Likely to Be Reported When Firm *j* Is Small? In this subsection, we write an equation for $\Pr[i \to j \text{ observed} \mid i \to j, X_{ij}]$ in terms of ϕ and probabilities that can be estimated from observable data. We will be estimating the probability that s lies in different subintervals of the [0, 1] interval using a multinomial logit regression. One of the dependent variables in the multinomial regression is the size of the customer. Setting all other covariates to their average value, we will allow the customer size to vary. This will allow us to determine the extent to which $\Pr[i \to j \text{ observed} \mid i \to j]$ depends on the size of the customer.

Again, by Bayes' Rule (Eq. **S13**),

$$\Pr[s_{ij} = s | i \to j] = \frac{\Pr[\textit{observed } i \to j \mid i \to j] \Pr[s_{ij} = s | \textit{observed } i \to j]}{\Pr[\textit{observed } i \to j \mid s_{ij} = s]}. \qquad \textbf{[S13]}$$

Because $1 = \int_0^1 \Pr[s_{ij} = s \mid X_{ij}, i \to j] ds$, we have (Eq. **S14**)

$$\int_0^1 \frac{\Pr[\textit{observed } i \to j \mid i \to j, X_{ij}] \Pr[s_{ij} = s \mid \textit{observed } i \to j, X_{ij}]}{\Pr[\textit{observed } i \to j \mid s_{ij} = s, X_{ij}]} ds = 1. \qquad \textbf{[S14]}$$

This implies (Eq. **S15**)

$$\frac{1}{\Pr[\textit{observed } i \to j | i \to j, X_{ij}]} = \int_{0.1}^1 \Pr[s_{ij} = s | \textit{observed } i \to j; X_{ij}] ds + \int_0^{0.1} \frac{\Pr[s_{ij} = s | \textit{observed } i \to j; X_{ij}]}{\phi(s)} ds. \qquad \textbf{[S15]}$$

Thus (Eq. **S16**),

$$\frac{1}{\Pr[\textit{observed } i \to j \mid i \to j, X_{ij}]} = \Pr[s_{ij} \geq 0.1 \mid \textit{observed } i \to j, X_{ij}] + \int_0^{0.1} \frac{\Pr[s_{ij} = s \mid \textit{observed } i \to j; X_{ij}]}{\phi(s)} ds. \qquad \textbf{[S16]}$$

We approximate the above equation by binning s in the following way (Eq. **S17**):

$$\frac{1}{\Pr[\textit{observed } i \to j | i \to j, X_{ij}]} \approx \Pr[s_{ij} \geq 0.1 | \textit{observed } i \to j, X_{ij}] + \sum_{k=1}^{n} \frac{\Pr\left[s_{ij} \in \left(\frac{1}{10}\frac{k-1}{n}, \frac{1}{10}\frac{k}{n}\right] \mid \textit{observed } i \to j, X_{ij}\right]}{\phi\left(\frac{1}{10}\frac{2k-1}{2n}\right)}. \qquad \textbf{[S17]}$$

With this approximation (Eq. **S18**),

$$\Pr[i\to j \text{ observed} \mid i\to j, X_{ij}] \approx \left(\Pr[s_{ij}\geq 0.1 \mid \textit{observed } i\to j, \mathrm{X}_{ij}] + \sum_{k=1}^{n} \frac{\Pr\left[s_{ij}\in\left(\frac{1}{10}\frac{k-1}{n}, \frac{1}{10}\frac{k}{n}\right] \mid \textit{observed } i\to j, X_{ij}\right]}{\phi\left(\frac{1}{10}\frac{2k-1}{2n}\right)} \right)^{-1}. \quad \textbf{[S18]}$$

Notice that if $\phi(s)$ were equal to 1 (edges are always reported) in the interval [0, 0.1], we would get that $\Pr[\textit{observed} i\to j \mid i\to j, X_{ij}] = (\Pr[S_{ij}\geq 0.1 \mid \textit{observed } i\to j, \mathrm{X}_{ij}] + \Pr[S_{ij}<0.1 \mid \textit{observed } i\to j, \mathrm{X}_{ij}])^{-1}$, which would imply that $\Pr[\textit{observed } i\to j \mid i\to j, X_{ij}] = 1$, as expected. Also note, as ϕ decreases, the second term on the right-hand side increases, and therefore, $\Pr[\textit{observed } i\to j \mid i\to j, X_{ij}]$ decreases.

Because we have an estimate of ϕ from the previous subsection, we can use Eq. **S18** to estimate $\Pr[i\to j \text{ observed} \mid i\to j, X_{ij}]$.

To calculate $\Pr[i\to j \text{ observed} \mid i\to j, X_{ij}]$, we need to estimate $\Pr[S_{ij}\geq 0.1 \mid \textit{observed } i\to j, \mathrm{X}_{ij}]$ and $\Pr[s_{ij}\in(\frac{1}{10}\frac{k-1}{n}, \frac{1}{10}\frac{k}{n}] \mid \textit{observed } i\to j, X_{ij}]$ for $k \in \{1, \ldots n\}$. We do so using a multinomial logit regression, with $n = 10$ bins. The coefficient estimates of such a regression are presented in Table S7. Observed edges are significantly more likely to be small when the customer is small and the supplier is large. The first relationship is expected: small customers are likely to demand small fractions of their suppliers' total sales.

Using the results of this regression and our estimate of ϕ, we are now in a position to estimate how the probability of observing an edge varies as the size of the customer varies. In particular, we consider how the probability varies for a hypothetical edge with the following characteristics:

i) the supplier and customer are both in the manufacturing sector (the modal sector),
ii) the year is set to 2000,
iii) the distance between the customer and supplier is in the (500, 1,000)-mi interval,
iv) log employment of the supplier is set to the mean value in our sample of observed edges, and
v) log employment of the customer is set to three different values: mean − SD, mean, and mean + SD. The SD of the log number of employees of the customer in the sample of observed edges is 2.05.

From the ϕ_2 column of Table S6 and the predicted probabilities from the multinomial logit regression, we are able to compute $\frac{\Pr[s_{ij}\in(\frac{1}{10}\frac{k-1}{n}, \frac{1}{10}\frac{k}{n}] \mid \textit{observed } i\to j, X_{ij}]}{\phi(\frac{1}{10}\frac{2k-1}{2n})}$ (Eq. **S18**).

Using Eq. **S18**, we compute (Eqs. **S19–S21**)

$$\Pr[i \to j \text{ observed} \mid i \to j, \text{big customer}] \approx 64\%, \quad \textbf{[S19]}$$

$$\Pr[i \to j \text{ observed} \mid i \to j, \text{mean customer}] \approx 57\%, \text{ and} \quad \textbf{[S20]}$$

$$\Pr[i \to j \text{ observed} \mid i \to j, \text{small customer}] \approx 50\%. \quad \textbf{[S21]}$$

From Eqs. **S19** and **S21** (Eq. **S22**),

$$\frac{\Pr[i \to j \text{ observed} \mid i \to j, \text{big customer}]}{\Pr[i \to j \text{ observed} \mid i \to j, \text{small customer}]} = 1.29. \quad \textbf{[S22]}$$

To conclude, we assess how the difference in undercounting of in-degree for small vs. large firms qualitatively affects the in-degree distribution. For every firm in 2006, we divide its in-degree by the probability that an actual edge is observed (giving an estimate of the actual number of suppliers for the firm). In Fig. S4, we compare the shape of the observed in-degree distribution with the estimated actual in-degree distribution. Qualitatively, the shape of the in-degree distribution is similar, whether we include the unobserved edges or not.

1. Saldana J (2007) Continuum formalism for modeling growing networks with deletion of nodes. *Phys Rev E Stat Nonlin Soft Matter Phys* 75:027102.
2. Risken H (1996) *The Fokker-Planck equation: Methods of solution and applications* (Springer-Verlag, Berlin, Germany).
3. Gabais X (2009) Power laws in economics and finance. *Annu Rev Econom* 1:255–294.
4. Sinko JW, Streifer W (1967) A new model for age-size structure of a population. *Ecology* 48:910–918.
5. Cohen L, Frazzini A (2008) Economic links and predictable returns. *J Finance* 63: 1977–2011.
6. Anderson JE, van Wincoop E (2004) Trade costs. *J Econ Lit* 42:691–751.
7. Disdier AC, Head K (2008) The puzzling persistence of the distance effect on bilateral trade. *Rev Econ Stat* 90:37–48.
8. Hortaçsu A, Martinez-Jerez FA, Douglas J (2009) The geography trade on eBay and MercadoLibre. *Am Econ J Micro* 1:53–74.
9. Currarini S, Jackson MO, Pin P (2010) An economic model of friendship: Homophily, minorities, and segregation. *Econometrica* 77:1003–1045.
10. Jackson MO, Rogers BW (2007) Meeting strangers and friends of friends: How random are social networks? *Am Econ Rev* 97:890–915.

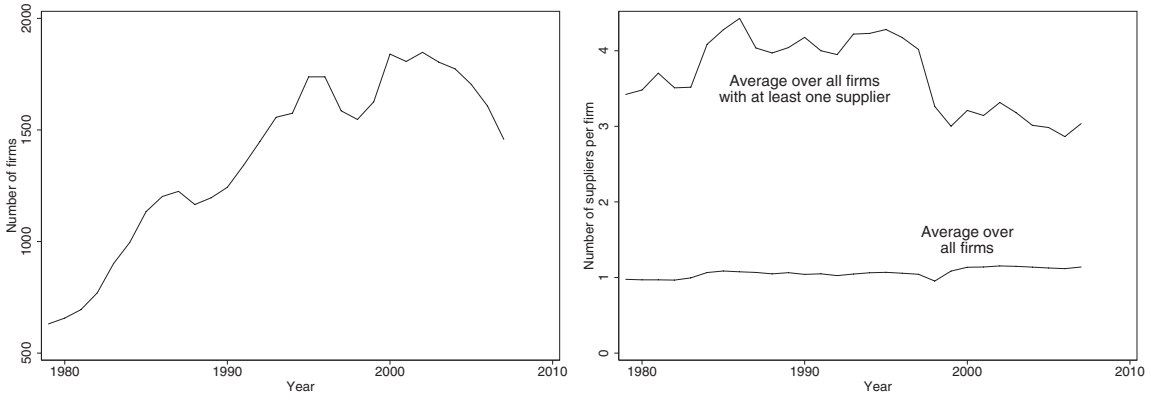

Fig. S1. (*Left*) Number of firms in the network. (*Right*) Average number of customers per firm.

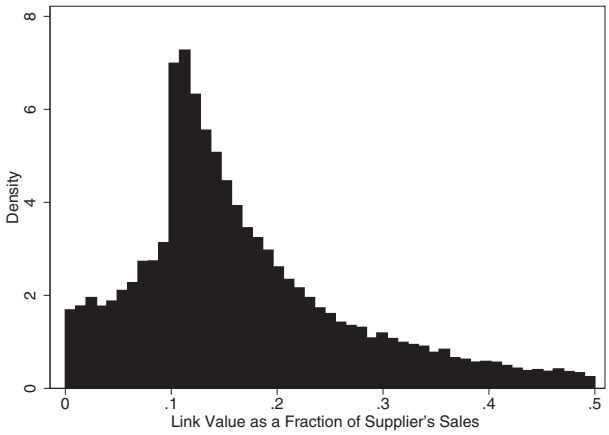

Fig. S2. Ratio of value of the edge to supplier's total sales.

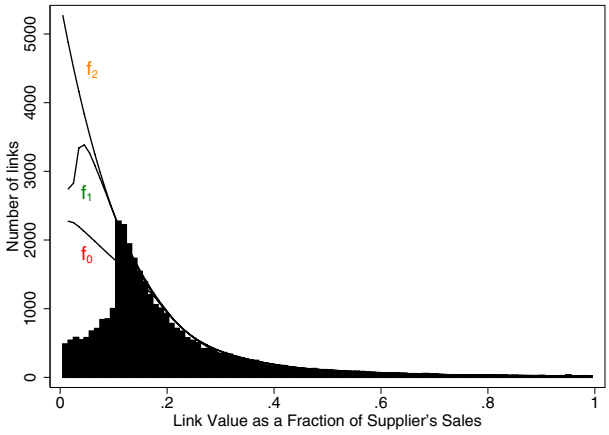

Fig. S3. Extrapolations using a local polynomial smoother on the number of edges that exist for $s < 0.1$. f_m is an mth-degree local polynomial smoother.

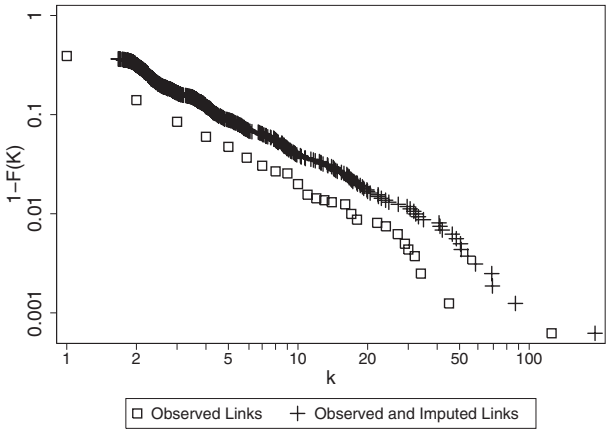

Fig. S4. Observed in-degree distribution for 2006 (squares) as well as the predicted in-degree distribution when missing edges are included (+ signs).

Table S1. Two-digit SIC industries with the largest total revenues (in billions of dollars) in our sample using data from 1997

SIC	Industry	Firms	Total revenues
37	Transportation equipment	65	625.2
48	Communications	73	403.6
29	Petroleum and coal products	13	392.8
35	Industrial machinery and equipment	151	341.4
28	Chemicals and allied products	170	303.9
53	General merchandise stores	12	270.2
73	Business services	241	251.4
36	Electrical and electronic equipment	198	247.0
50	Wholesale durable goods	36	201.4
20	Food and kindred products	35	175.5

Table S2. MLE estimates

Years	Existing → existing	New → existing	Pooled
1980–2007	0.169 (0.003)	0.181 (0.001)	0.176 (0.001)
1980–1989	0.125 (0.001)	0.170 (0.005)	0.151 (0.004)
1990–1999	0.108 (0.000)	0.168 (0.006)	0.142 (0.001)
2000–2007	0.140 (0.003)	0.208 (0.007)	0.172 (0.005)

Each cell is an MLE estimate of r using a different subsample of the dataset. For this table, we define $k_i(t-1)$ as the number of edges received by vertex i that is present in both years $t-1$ ant t.

Table S3. Coefficient estimates from logit regressions

Independent variable	Regression 1	Regression 2	Regression 3	Regression 4	Regression 5
Previous in-degree of supplier	0.003	−0.004	−0.018	−0.017	0.002
Previous in-degree of customer	0.037	0.028	0.027	0.027	0.193
Log employees of supplier		0.047	0.063	0.063	−0.098
Log employees of customer		0.634	0.860	0.858	0.483
Distance < 25 mi		1.193	0.844	0.838	0.610
Distance in (25, 100) mi		0.490	0.341	0.337	0.224
Distance in (500, 1,000) mi		−0.286	−0.209	−0.211	−0.154
Distance in (1,000, 1,500) mi		−0.518	−0.447	−0.445	−0.308
Distance in (1,500, 2,000) mi		−0.448	−0.414	−0.415	−0.400
Distance in (2,000, 2500) mi		−0.428	−0.425	−0.425	−0.343
Distance > 2,500 mi		−0.068	−0.203	−0.205	−0.173
Distance measure does not exist		−0.471	−0.599	−0.599	−0.357
Same one-digit SIC			−0.046	−0.046	−0.109
Same two-digit SIC			1.290	1.287	0.939
Same three-digit SIC			2.316	2.309	1.557
Same four-digit SIC			2.611	2.608	2.105
Labor productivity of supplier			0.087	0.089	0.061
Labor productivity of customer			0.970	0.966	0.513
Number of common edges				0.828	0.638
Did the edge exist in the previous year?					8.693
Did the edge exist 2 or 3 y ago?					4.849
Did the edge exist 4+ y ago?					4.164
N	7,817,891	6,784,360	6,758,319	6,758,319	6,758,319
Pseudo-R^2	0.071	0.100	0.148	0.148	0.611

The dependent variable is the probability that an edge exists between vertex i and vertex j in a particular year. Year-level fixed effects are included. Errors are clustered by ij pair. In the fourth column, all coefficients—except for previous in-degree of supplier and same one-digit SIC—are statistically significant at the 1% level.

Table S4. Coefficient estimates of logit regressions

Independent variable	Regression 1	Regression 2	Regression 3	Regression 4
Previous in-degree of customer	0.032	0.025	0.024	0.023
Log employees of supplier		0.028	0.058	0.058
Log employees of customer		0.512	0.700	0.700
Distance < 25 mi		1.102	0.816	0.815
Distance in (25, 100) mi		0.277	0.160	0.159
Distance in (500, 1,000) mi		−0.316	−0.242	−0.243
Distance in (1,000, 1,500) mi		−0.504	−0.432	−0.432
Distance in (1,500, 2,000) mi		−0.503	−0.455	−0.455
Distance in (2,000, 2,500) mi		−0.430	−0.395	−0.395
Distance > 2,500 mi		0.042	−0.071	−0.071
Distance measure does not exist		−0.352	−0.452	−0.452
Same one-digit SIC			−0.094	−0.094
Same two-digit SIC			1.194	1.194
Same three-digit SIC			2.082	2.081
Same four-digit SIC			2.658	2.656
Labor productivity of supplier			0.112	0.112
Labor productivity of customer			0.761	0.760
Number of common edges				0.490
N	2,037,749	1,639,491	1,633,554	1,633,554
Pseudo-R^2	0.037	0.059	0.103	0.103

The dependent variable is the probability that an edge exists between vertex *i* and vertex *j* in a particular year conditional on vertex *i* entering the network. Year-level fixed effects are included. Errors are clustered by *ij* pair. In the last column, all coefficients—except for common links, same one-digit SIC, distance in (25, 100) mi, and distance >2,500 mi—are statistically significant at the 1% level.

Table S5. Coefficient estimates and robust SEs from a logit regression

	β	SE
Year	0.003	0.007
Log employment customer	0.048	0.060
Log employment supplier	−0.057	0.049
Log assets supplier	0.034	0.048
Log assets customer	−0.056	0.057
Same industry	−0.052	0.126
Supplier industry		
Construction	0.010	0.497
Manufacturing	0.289	0.246
Transportation	−0.128	0.274
Wholesale	−0.067	0.335
Retail	−0.714	0.516
FIRE	−0.570	0.383
Services	0.211	0.259
Public administration	−0.791	0.431
Customer industry		
Construction	0.674	1.151
Manufacturing	−0.033	0.335
Transportation	−0.039	0.353
Wholesale	−0.096	0.391
Retail	0.005	0.394
FIRE	0.319	0.461
Services	0.137	0.382
Public administration	0.227	0.566
Distance < 25 mi	−0.174	0.181
Distance in (25, 100) mi	−0.398	0.181
Distance in (100, 500) mi	−0.057	0.135
Distance in (1,000, 1,500) mi	−0.160	0.143
Distance in (1,500, 2,000) mi	−0.088	0.183
Distance in (2,000, 2,500) mi	0.087	0.207
Distance > 2,500 mi	0.539	0.259
Distance measure does not exist	0.076	0.189
Constant	−5.157	13.679

The dependent variable is the probability that $s_{i,j}$ is less than 0.10 conditional on being between 0.09 and 0.11. FIRE, finance, insurance, and real estate.

Table S6. Number of edges and fraction of existing edges that are observed in the dataset

Interval	Observed edges	f_0	f_1	f_2	φ_0	φ_1	φ_2
(0, 0.01)	487	—	—	5,269	—	—	0.092
(0.01, 0.02)	541	2,273	2,743	4,876	0.238	0.197	0.111
(0.02, 0.03)	582	2,253	2,828	4,509	0.258	0.206	0.129
(0.03, 0.04)	548	2,194	3,341	4,162	0.250	0.164	0.132
(0.04, 0.05)	579	2,125	3,385	3,836	0.272	0.171	0.151
(0.05, 0.06)	675	2,053	3,268	3,532	0.329	0.207	0.191
(0.06, 0.07)	712	1,981	3,084	3,247	0.356	0.231	0.219
(0.07, 0.08)	841	1,907	2,892	2,980	0.441	0.291	0.282
(0.08, 0.09)	852	1,833	2,693	2,732	0.465	0.316	0.312
(0.09, 0.1)	1,002	1,761	2,489	2,501	0.567	0.403	0.401

Table S7. Coefficient estimates from a multinomial logit regression

Omitted alternative: $s \geq 0.1$	(0, 0.01)	(0.01, 0.02)	(0.02, 0.03)	(0.03, 0.04)	(0.04, 0.05)
Year	0.001	0.020	−0.001	−0.013	0.005
Log employment customer	−0.298	−0.232	−0.211	−0.209	−0.141
Log employment supplier	0.337	0.274	0.230	0.244	0.135
Same industry	−0.104	−0.256	−0.338	0.010	−0.006
Supplier industry					
Agriculture	1.562	−0.204	1.499	0.875	−14.985
Construction	0.796	0.719	1.154	1.245	0.594
Manufacturing	−0.018	−0.066	0.208	−0.085	−0.215
Transportation	1.025	0.483	0.268	0.159	0.509
Wholesale	0.455	−0.163	0.495	0.169	0.294
Retail	2.211	1.272	0.542	1.192	0.354
FIRE	3.918	3.264	3.201	2.329	1.519
Services	0.551	0.112	0.028	0.150	0.227
Public administration	0.013	1.193	0.736	0.807	0.167
Customer industry					
Agriculture	−14.695	−15.716	−15.721	−15.031	−15.571
Construction	0.360	−16.585	−16.669	−0.615	−0.614
Manufacturing	0.674	−0.078	0.059	0.469	0.078
Transportation	0.356	−0.414	−0.511	−0.170	−0.280
Wholesale	−0.248	−2.170	−1.399	−1.219	−1.134
Retail	0.310	−0.652	−0.503	−0.068	−0.169
FIRE	−1.132	−1.507	−1.930	−1.037	−0.923
Services	0.479	−0.646	−0.689	0.356	−0.023
Public administration	1.210	−0.895	−0.377	0.110	0.038
Distance < 25 mi	0.666	0.345	−0.477	−0.066	−0.470
Distance in (25, 100) mi	0.278	−0.072	−1.067	−0.135	−0.369
Distance in (100, 500) mi	0.151	0.471	−0.061	0.275	0.084
Distance in (1,000, 1,500) mi	0.166	0.331	−0.111	−0.348	−0.198
Distance in (1,500, 2,000) mi	−0.183	0.135	−0.401	−0.261	−0.071
Distance in (2,000, 2,500) mi	0.238	0.483	0.108	−0.237	−0.180
Distance > 2,500 mi	0.105	−0.070	0.001	−0.321	−0.606
Distance measure does not exist	0.280	0.699	−0.057	0.345	0.158
Constant	−6.029	−42.193	0.070	23.578	−13.072

Omitted alternative: $s \geq 0.1$	(0.05, 0.06)	(0.06, 0.07)	(0.07, 0.08)	(0.08, 0.09)	(0.09, 0.1)
Year	0.023	0.011	0.006	0.009	0.005
Log employment customer	−0.117	−0.130	−0.053	−0.083	−0.062
Log employment supplier	0.158	0.140	0.147	0.106	0.150
Same industry	0.020	0.131	0.004	−0.108	−0.080
Supplier industry					
Agriculture	−0.306	−0.486	−15.032	−0.456	−15.102
Construction	0.185	0.608	0.031	0.014	0.096
Manufacturing	0.275	0.092	−0.223	−0.306	−0.288
Transportation	0.500	0.355	0.116	0.185	−0.015
Wholesale	0.814	0.268	−0.214	0.076	0.067
Retail	0.599	−1.045	−0.400	0.207	0.623

Table S7. Cont.

Omitted alternative: $s \geq 0.1$	(0.05, 0.06)	(0.06, 0.07)	(0.07, 0.08)	(0.08, 0.09)	(0.09, 0.1)
FIRE	1.531	1.085	0.220	0.038	0.370
Services	0.537	0.296	0.020	−0.156	−0.190
Public administration	0.946	1.257	0.290	0.709	0.625
Customer industry					
Agriculture	−15.466	−15.657	−15.346	−15.226	−15.649
Construction	−0.456	−15.695	−15.305	−15.041	−0.559
Manufacturing	−0.045	−0.147	0.058	0.282	−0.284
Transportation	−0.043	−0.513	−0.298	0.066	−0.409
Wholesale	−1.142	−0.824	−0.600	−0.220	−0.340
Retail	−0.100	−0.282	−0.280	0.133	−0.418
FIRE	−1.051	−0.675	−0.101	−0.149	−0.619
Services	−0.229	−0.171	0.060	0.144	−0.549
Public administration	0.326	−1.200	0.061	0.550	−0.709
Distance < 25 mi	−0.075	0.002	0.009	−0.034	−0.159
Distance in (25, 100) mi	−0.369	−0.104	−0.233	0.227	0.274
Distance in (100, 500) mi	0.301	0.541	0.195	0.133	0.001
Distance in (1,000, 1,500) mi	−0.161	0.317	0.116	0.355	0.071
Distance in (1,500, 2,000) mi	0.175	0.355	0.087	0.365	0.042
Distance in (2,000, 2,500) mi	0.479	0.362	0.381	−0.033	−0.130
Distance > 2,500 mi	−0.030	0.099	−0.044	−0.094	−0.455
Distance measure does not exist	0.338	0.322	0.066	0.221	−0.217
Constant	−48.798	−25.898	−14.308	−21.781	−12.876

The dependent variable is the probability that s falls in a particular bin, with the $s > 0.1$ bin being the omitted alternative.