Decision Support Systems 99 (2017) 9–17

Contents lists available at ScienceDirect

Decision Support Systems

journal homepage: www.elsevier.com/locate/dss

Geography of online network ties: A predictive modelling approach

Swanand J. Deodhar *, Mani Subramani, Akbar Zaheer

Carlson School of Management, University of Minnesota, 321 19th Avenue South, Minneapolis, MN 55455, United States

ARTICLE INFO

Article history:
Received 20 September 2016
Received in revised form 24 April 2017
Accepted 4 May 2017
Available online 6 May 2017

Keywords:
Geography
Online network ties
Psychic distance
Predictive modelling
Cluster analysis

ABSTRACT

Internet platforms are increasingly enabling individuals to access and interact with a wider, globally dispersed group of peers. The promise of these platforms is that the geographic distance is no longer a barrier to forming network ties. However, whether these platforms truly alleviate the influence of geographic distance remains unexplored. In this study, we examine the role of geographic distance with machine learning approach using a unique dataset of the network ties between traders in an online social trading platform. Specifically, we determine the extent to which, compared to other types of distances, geographic distance predicts the occurrences of the network ties in country dyads. Using cluster analysis and predictive modelling, we show that not only the geographic distance and network ties exhibit an inverse association but also that geographic distance is the strongest predictor of such ties.

1. Motivation for the research

Since the inception of internet, the world is moving towards a scenario in which individuals, irrespective of their location, can interact with a large, globally dispersed networks of peers, leading to transformative economic activities such as user innovation [26], online labor markets [9], and crowdsourcing [22]. Internet platforms drive this transformation by providing means to access and communicate with one's peers at practically non-existent cost. For such online intermediaries, the ideal world is the one in which individuals can engage in frictionless interactions and create network ties, rendering geographic distance inconsequential [11]. Although the idea of frictionless interactions is appealing, there is surprisingly little prior empirical research that informs the fundamental question – what is the role geographic distance in shaping online behavior?

A few recent studies have examined this issue [10] suggesting that distance adversely influences frequency and magnitude of dyadic ties in online context as well. Our study extends this literature in at least two ways. First, we isolate the extent to which geographic distance predicts the occurrences of dyadic ties by comparing its predictive power with that of the competing distance measures. Such an approach is necessary because a standard online platform does not directly provide geographic distance as an information cue to its users. Instead, it makes each user's nationality and other location information visible to other users. Therefore, geographic distance is only one of the several distance measures that can predict user's behavioral response. Any assessment of geographic distance as a predictor of network ties in an online context is incomplete without the inclusion of other forms of distances. Second, we examine geographic distance in a context that does not follow the two-sided platform structure which is predominant in the extant literature on distance effects in online settings. This difference is relevant to the occurrence of dyadic ties because on two-sided platforms dyadic ties are typically “cross-sided.” Instead, our setting allows any user to form a tie with any other user on the platform, broadening the possible pool of users with whom ties can be established. In sum, our primary research question is as follows: *in a globally distributed network of individuals, in which users can create ties with any other user, whether and to what extent geographic distance predicts the occurrence of dyadic ties*?

We address this research question by using a dataset of dyadic ties obtained from an electronic investment platform, which we refer to as XTrader. The platform is meant for the currency and commodities trading and has a user-base of over a million traders, representing nearly 100 countries. XTrader is an appropriate choice for our study for several reasons. First, the platform allows traders to form direct ties with each other. Because all the traders are engaged in the same activity (i.e. trading), there are no distinct sides to the platform. Hence, every trader can form a tie with every other trader. Second, a trader can create a tie only by allocating a certain portion of their fund to the other trader. That is, each tie that a trader creates has a cost associated with it, allowing us to consider the existence of a tie as a conscious decision on a trader's part for which the trader is likely to consider available information cues about a potential tie partner. Third, the platform provides each trader's country as the only demographic information cue. This cue is publicly visible to everyone. The salience of trader's nationality enables the distance mechanism to come into play.

* Corresponding author.
E-mail addresses: deodh009@umn.edu (S.J. Deodhar), subra010@umn.edu (M. Subramani), azaheer@umn.edu (A. Zaheer).

http://dx.doi.org/10.1016/j.dss.2017.05.010
0167-9236/© 2017 Elsevier B.V. All rights reserved.

Table 1
Summary of articles on distance effect in online contexts.

Study	Distance measure	Context	Key finding related to distance effect
Blum & Goldfarb [6]	1. Geographic distance, 2. Difference in GDP	Website visits	In taste-based digital products (e.g. music, games) users from a given country make significantly more visits to websites from geographically nearer countries.
Gefen & Carmel [18]	1. Geographic distance	Online labor markets (Rent A coder)	Except for the American clients, others tend to award project a geographically proximal agent. However, the preference gets mitigated if the agent and the client come from English-speaking countries.
Hortaçsu, Martínez-Jerez, & Douglas [21]	1. Geographic distance	Online auction sites (MercadoLibre and eBay)	Buyers prefer to buy from sellers who are geographically proximal, preferably, within a driving distance or same city.
Agarwal, Catalini, & Goldfarb [3]	1. Geographic distance	Crowdfunding (Sellaband)	Geographically proximal investors are likely to fund a project early while the distant ones invest once the project accumulates investments
Takhteyev, Gruzd, & Wellman [41]	1. Geographic distance 2. Language similarity 3. Air travel (number of direct flights between two locations)	Microblogging (Twitter)	Geographic distance has an adverse influence on the number of Twitter ties. However, it is not robust to the magnitude of air travel between two locations.
Burtch, Ghose, & Wattal [10]	1. Geographic distance 2. Cultural distance	Crowdfunding (Kiva.com)	Both geographic and cultural distance significantly reduce the number of ties between donors and receivers The two distances are substitutes of each other
Lin & Viswanathan [27]	1. Geographic distance	Crowdfunding (Prosper.com)	Donors are more likely to donate money to those who are geographically proximal even when such a donation is not in the donor's economic interest
Posegga, Zylka, & Fischbach [32]	1. Geographic distance	Crowdfunding (Kickstarter)	While there exists gender-based (male(female) investors investing money in projects started by male(female) entrepreneurs), geographic distance does not matter.
Lengyel et al. [25]	1. Geographic distance between towns	Intra-Country Social Network in Hungary (International Who is Who)	Individuals are more likely to form links with those who are from geographically proximal towns. However, in the online social network, the probability of a tie's existence decays at a slower rate with respect to distance.

We adopt the data construction similar to Burtch et al. [10]. The dataset we obtained from XTrader pertains to trader-dyads observed for 46 consecutive weeks. Because we are interested in predicting the number of ties between pairs of countries, we begin by aggregating these observations to country-dyads. In the final dataset, we have 266,570 data points consisting of 5795 unique pairs observed over 46 weeks. These pairs represent 95 distinct countries. In each country-dyad, the outcome variable is the count of distinct ties from one country to another, and the predictors are different types of distances between the countries in that particular dyad. More specifically, we use psychic [13] and geographic distance measures.

We employ cluster analysis and predictive modelling because we are interested in assessing the extent to which distances and specifically, geographic distance predicts the occurrences of ties. First, using cluster analysis, we obtain a robust, 4-cluster solution. Among the 4 clusters, one cluster represents country dyads that have the highest geographic

Table 2
Variable descriptions and summary statistics.

Variable name	Brief description	Source	Mean	Standard deviation
Count of ties in Country Dyads	The number of unique ties from the source country to the destination country	Obtained from the platform	237.24	176.37
Geographic distance	The distance (in KM) between two most populated agglomerations. The distance is computed using the latitudes and longitudes of the agglomerations	Mayer & Zignago [28]	7216.27	4831.15
Social (political)[a]	The difference in the political ideology scale for the two countries	Dow & Karunaratna [14]	0.41	0.27
Democracy (political)[a]	The composite score that captures the difference between two countries on a series of scales including POLICON, Political rights scale, civil liberty scale, and POLITY V scale.	Dow [13]	0.56	0.48
Religion	A composite score capturing the extent to which: • The major religions of the two countries are different • One country practices the religion of the other country		0.79	0.44
Language	A composite score capturing the extent to which: • The national languages of the two countries are different • The population in one country speaks the national language of the other country		1.51	1.36
Education	The composite score for the • The difference in the literate adults • Population enrolled in the second as well as the third levels of education		0.79	0.58
Industrial Development	The difference in industrial development between two countries. Dow [13] measures industrial development using ten sub-factors • GDP difference • Ownership of electronic goods (TV, radio, telephones) and car • Energy consumption • Newspaper usage • Percentage of the URBAN population		0.82	0.58

[a] For the original references of each scale, see Dow [13].

Table 3
Centroid table.

Cluster	Copy count	Geographic distance (in KM)	Industrial development	Social (political)	Democracy (political)	Religion	Language	Education	Observations
Cluster-1	270.570	2036.888	0.595	0.356	0.116	0.845	0.338	0.590	729
Cluster-2	3.135	2364.858	0.500	0.400	0.424	0.892	0.672	0.465	92,745
Cluster-3	1.804	7740.023	1.464	0.482	0.902	1.002	0.535	1.394	76,990
Cluster-4	***0.761***	***11,838.620***	***0.573***	***0.420***	***0.461***	***0.972***	***0.496***	***0.575***	***96,106***

distance and the lowest number of ties, suggesting that at the country level, greater geographic distance is associated with fewer ties. Next, using predictive modelling, we find that the geographic distance is the strongest predictor of the number of ties between a country-dyad. In particular, among all the distance measure, exclusion of geographic distance from predictive models leads to the highest increase in root mean square error (RMSE). In sum, our findings clearly indicate that when presented with other users' nationalities as an information cue, the geographic distance most accurately predicts the creation of network ties. More broadly, our study shows that geographic distance-based preference is evident even in contexts that allow a focal user to form ties with any other user on the platform, negating the death of distance hypothesis.

The paper is structured as follows: in the next section, we review the background literature on the role of distance in online contexts. Next, we provide an overview of the empirical context as well as the variables of interest. Then, we discuss the analysis approach, followed by cluster analysis and predictive modelling results. We demonstrate the stability of the findings using multiple robustness checks. In the last section, we summarize the study's contribution and suggest directions for future research.

2. Background literature: distance effect in online contexts

In dyadic exchanges occurring in offline settings, the influence of distance is well documented [39]. However, the topic has recently become popular in studies of online settings as well. The role of distance in online contexts revolves around the death of distance hypothesis which states that with the emergence of the internet and the resulting reduction in communication and information costs, distance no longer acts as a barrier to social and economic exchanges [11]. However, recent studies have found some contrary evidence (Table 1). For instance, Burtch et al. [10] found that on an online pro-social lending platform, geographic and cultural distances between the donor and the recipient's respective countries reduce the lending activity. Similarly, Lin & Viswanathan [27] show that donors are likely to give money to those who are geographically proximal even when such lending has no economic rationale. Gefen & Carmel [18] report similar findings in their study of online labor markets. They show that except the American clients, others prefer to award projects to geographically proximal agents. In general, these studies indicate that even with the emergence of the internet, the world is far from flat.

While there is consistency in the findings of recent studies, there are at least two research gaps. First, existing studies only incorporate geographic distance as a single measure of distance, excluding other possible forms of distance which could equally predict the occurrence of ties. Indeed, the extant literature on offline contexts suggests that geographic distance is only one of the several distance measures [5,14]. We argue that because online platforms typically provide the nationality of each user as an information cue, the choice of forming a tie could, therefore, also be associated with other types of distances that exist between two user's countries. To accurately assess the extent to which geographic distance predicts network ties, one needs to compare its predictive power with that of the alternative distance measures as well. We address this gap by adopting a set of psychic distance measures [13,14]. Psychic distance, a multi-dimensional construct measuring distance between countries, is defined as the sum of individuals' perceptions regarding the differences between two countries regarding a variety of dimensions such as language, education, and culture ([38]; p. 832). Along with geographic distance, it is one of the most commonly studied forms of distance [30].

Second, current understanding of distance effects on online tie occurrences is mostly limited to two-sided platforms. On such platforms, the possible pool of users with whom a focal user can form a tie is already constrained by the market sides. For example, on crowdfunding platforms, only the ties between the investors and the recipients can exist. Similarly, in online labor markets, the ties are between the service provider and the client. However, all online platforms do not follow the two-sided structure. For instance, in online communities, any user can form a tie with any other user. In conclusion, on some online platforms, the existence of a tie is a function of the platform's two-sided nature while no such constraint exists in online platforms that do not have any explicitly identified sides. Given this fundamental difference and its implications for dyadic ties, it is important to separately examine to what extent distances predict ties in an online setting that does not follow a two-sided market structure. We address this gap by adopting an empirical context of an online trader networks in which any trader can form a tie with any other trader.

Our choice of psychic distance as the set of other distance measures is meaningful. Psychic distance influence occurrences of network tie given its fundamental operating mechanism: it creates a “disturbance in information flows arising out of actors' perception” ([30]; p. 827). Larger the psychic distance between two entities, more difficult it may be to obtain and interpret the information about each other. As a result, at least in offline settings, “larger the psychic distance, more difficult it is to build new relationships” ([23]; p.1414) while “short psychic distance will facilitate the establishment and development of relationships” ([23]; p.1426). A similar argument can be made about ties in online networks. In our context, the occurrences of ties between country dyads could be associated with psychic distance because traders may find it difficult to interpret the information (e.g. investment decisions) about traders from a country that is “psychically far.”

Table 4
Centroid table after excluding traders from most represented countries.

Cluster	Copy count	Geographic distance (in KM)	Industrial development	Social (political)	Democracy (political)	Religion	Language	Education	Observations
Cluster-1	51.218	5207.401	0.906	0.347	0.290	1.080	0.590	0.952	1875
Cluster-2	1.122	2566.802	0.510	0.405	0.468	0.902	0.701	0.501	82,482
Cluster-3	0.665	7981.966	1.467	0.475	0.912	1.009	0.539	1.371	65,357
Cluster-4	***0.403***	***11,885.048***	***0.565***	***0.427***	***0.464***	***0.973***	***0.479***	***0.573***	***89,624***

Table 5
Centroid table after adding the count of traders from each country.

Cluster	Copy Count	Geographic distance (in KM)	Investors count (source)	Investor count (destination)	Industrial development	Social (political)	Democracy (political)	Religion	Language	Education	Observations
Cluster 1	30.5139	5283.991	3968.542	24,062.750	0.918	0.4397	0.4485	0.933	0.585	0.839	11,786
Cluster 2	10.1306	5441.484	24,228.21	2673.369	0.924	0.4424	0.4779	0.938	0.585	0.841	12,046
Cluster 3	1.3268	2548.463	2434.565	2508.971	0.570	0.4067	0.5052	0.902	0.680	0.573	90,997
Cluster 4	***0.6087***	***10,509.970***	***2258.449***	***2292.849***	***0.927***	***0.4440***	***0.6342***	***0.985***	***0.498***	***0.883***	***1,51,741***

3. Empirical context and variable descriptions

3.1. Empirical context

Over the last few years, financial investments and trading industry has seen a slew of technological innovations that have transformed the business practices. The incorporation of social media technologies into retail trading is one such advancement [37]. Social media features let the investors transparently view the real-time trading data-feeds of other investors on the platform, allowing them to draw on the skills of a globally dispersed network of peers. In this study, we use a dataset obtained from one such platform, which we refer to as XTrader. It is one of the leading platforms that combine investments with social media. XTrader offers forex and stock trading facilities to retail traders in nearly 100 countries around the world. The process of participating on XTrader is as follows: traders wanting to trade are required to open an account with a certain minimum amount in their local currencies. Once XTrader verifies their bank details and links their bank account to their platform profile, traders can begin to invest. The platform provides all the necessary information cues such as current market prices, company news feeds, and financial news providers.

However, the most salient feature of such platforms is copy-trading [17]. It is the primary mechanism through which traders can form ties with one another. The copy trading process on XTrader and other similar platforms work as follows: a trader can see the activities and information of other traders including their currently open as well as past investment positions, past losses and gains, and, most importantly, the country of origin. After assessing these cues, a focal trader *i* can allocate a certain portion of their fund to another trader *j*. Once the allocation takes place, the platform automatically replicates all the current as well as subsequent investments by *j* for *i*'s portfolio. In other words, *i* can grant a partial control of their own portfolio to *j*. The process is beneficial for both the traders. Because trader *i* can copy trade with multiple traders simultaneously, *i* can levelage the investment expertise of a larger pool of traders. On the other hand, trader *j* receives monetary reward from the platform if the number of traders copying *j* goes beyond a threshold.

The action of allocating funds creates a network of traders in which a tie between two traders represents a unidirectional allocation of monetary resources. For trader *i*, an outgoing tie to trader *j* $\{i - \rightarrow j\}$ implies that *i* has allocated funds to *j* while an incoming tie from trader *k* $\{k - \rightarrow i\}$ implies that *i* has received funds from *k*. Thus, any trader can have both incoming and outgoing ties at the same time. Moreover, given the economic nature of network tie and the risk which the traders run by relinquishing the control of their portfolio, it is possible to view the presence of a dyadic tie as a conscious decision on the trader's part. As a result, before allocating funds, a trader is likely to consider the available information cues about other traders. One such information cue, which is visible on each trader's profile, is their country of origin. We employ the salience of country of origin as a basis to assess the importance of distance.

3.2. Variable descriptions

3.2.1. Outcome variable

The study's primary outcome variable is the count of network ties for a given country-dyad. We compute this variable as follows: first, we aggregate the original dataset from trader-dyads to country dyads. For a particular country dyad $\{C_i - \rightarrow C_k\}$, observed in week t, we count the number of unique ties that the traders from country C_k receive from those from the country C_i. Note that a particular trader i from C_i might have ties with n distinct investors from C_k. In such a scenario, the resulting n ties are counted separately. Also, because the ties are unidirectional, number of ties for the country dyad $\{C_i - \rightarrow C_k\}$ is different from that for $\{C_k - \rightarrow C_i\}$. For example, in the 1st observation week, for the dyad {*Australia* — →*Spain*}, the number of distinct network ties was 96. However, in the same week, for the dyad {*Spain* — →*Australia*}, the outcome variable value was only 13. Finally, if country C_k received no incoming ties from traders in country C_i, we set the outcome variable to 0 for that particular dyad for that observation week. After aggregating the data from trader-dyads to country-dyads, we obtain 9, 025 unique dyads comprising of 95 countries. We observe each dyad for 46 consecutive weeks.

3.2.2. Predictor variables (geographic distance)

For geographic distance, we use Mayer & Zignago's [28] measure for the physical distance between countries. The distance is computed using the Great Circle Formula which uses the latitude and longitude information of the most populated agglomerations [28]. Our choice of the measure is in-line with the recent studies looking distance effect in online [10] as well as offline settings [39]. We obtain geographic distance in kilometers for all the 9025 country dyads.

3.2.3. Predictor variables (psychic distance)

We borrow additional distance measures from the extant literature on psychic distance [15,16]. Although several researchers have developed psychic distance measures (e.g. [8,20]), we adopt Dow's [13] measures primarily because of their wider coverage. The psychic distance

Table 6
Centroid table after excluding 50% randomly chosen observations.

Cluster	Copy count	Geographic distance (in KM)	Industrial development	Social (political)	Democracy (political)	Religion	Language	Education	Observations
Cluster 1	280.437	1990.725	0.583	0.361	0.116	0.842	0.343	0.581	350
Cluster 2	3.203	2358.832	0.497	0.398	0.423	0.891	0.660	0.460	46,158
Cluster 3	1.746	7725.636	1.460	0.483	0.902	1.003	0.547	1.389	38,912
Cluster 4	***0.772***	***11,842.450***	***0.572***	***0.421***	***0.461***	***0.973***	***0.494***	***0.575***	***47,865***

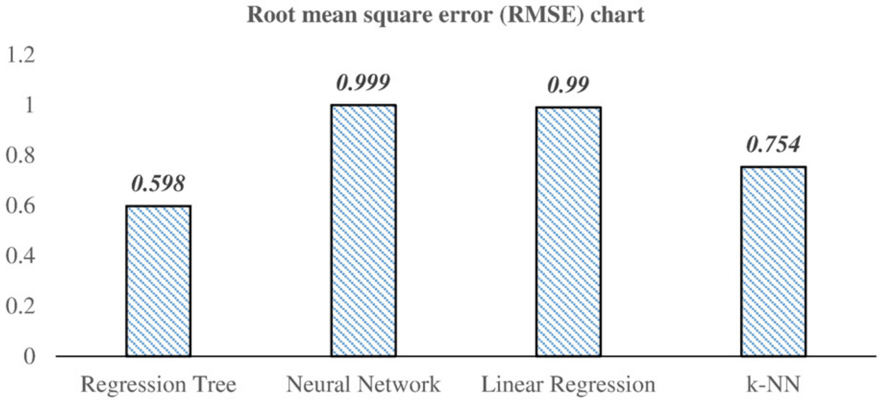

Fig. 1. Comparison of the predictive performance of the modelling techniques.

scores are available for each of the five dimensions (i.e. language, religion, education, industrial development, and political system) [13, 14].[1] Out of the 9025 country dyads, we do not have the distance scores for the political regime (subdivided into social and democracy) score for 3230 dyads (n = 148,580). Therefore, we exclude these pairs from any subsequent analysis. Thus, the final dataset has 266,570 observations consisting of 5795 unique country dyads. Table 2 summarizes the outcome and the predictor variables.

4. Analysis

4.1. Machine learning in information systems

In this study, our primary objective is to examine whether and to what extent does geographic distance predicts the occurrences of dyadic ties in a context which allows any user can form a tie with any other users. Accordingly, we adopt clustering and predictive modelling as the primary analysis techniques. In recent years, the role of machine learning techniques in information systems research has gained acceptance [35]. Current applications of predictive modelling in IS research include dynamic decision making [29], financial fraud detection [1], bidder strategies in online auctions [4], risk-based classification of software projects [36], and fake websites detection [2]. Explanatory and predictive models serve very different purposes [34]. Explanatory models are aimed at assessing the “strength of associations” between variables ([35]; p. 561). These models employ R^2 and *p*-values to assess the strength and the significance of estimated coefficients respectively. On the other hand, predictive modelling is concerned with “accurate predictions of observations that are either temporally or spatially new” ([35]; p. 555).[2] More formally, a predictive model is “the process of applying a statistical model or data mining algorithm to data for the purpose of predicting new or future observations.” ([34]; p. 291). Simply put, with predictive models, one cannot make a judgment of the causal influence of a unit change in the input variable on the output variable. Instead, one can assess whether a given input variable improves the accuracy of predicting future occurrences of the output variable.

4.2. Analysis process

We conduct the analysis in two steps. In the first step, we undertake cluster analysis using X-means algorithm [31]. We wish to assess whether country dyads can be categorized in latent groups based on distance measures and network ties. For such tasks, cluster analysis is a well-accepted approach. For example, Rai et al. [33] employed cluster analysis to identify four distinct patterns of adoption of electronic procurement innovations. Wallace, Keil, & Rai [36] used clustering to develop a risk-based classification of software products. Similarly, Bapna et al. [4] used clustering to identify archetypes of bidder strategies in online auctions. Given the wide variation in the numerical ranges of variables, we normalize each variable for cluster analysis.

The central mechanism underlying the X-means algorithm is as follows: the algorithm accepts an integer range. The upper and the lower bounds respectively indicate the maximum and the minimum number of centroids. The assumption is that the true count of centroids lies within this range. We provide the range of [2, 60].[3] Provision of a range is the key differentiator between the X-means and the *k*-means algorithms. In the latter approach, users have to pre-specify an exact number of centroids, increasing the chances of achieving only a “local optima” ([31]; p. 1), and therefore an inefficient solution. On the other hand, X-means uses information gain as a criterion to determine the number of centroids. The algorithm begins clustering by creating as many centroids as the lower bound and moves up incrementally. At each iteration, the algorithm computes the information criteria and uses it to determine whether more centroids should be calculated. It terminates either when adding new centroids does not lead to information gain ([31]; p. 3) or when the count of centroids reaches the upper bound.

In the second stage of analysis, we develop a series of predictive models to assess the extent to which geographic distance predicts occurrences of network ties. We accomplish this task in two stages. First, we build a set of models by including psychic and geographic distance measures. From this set, we select the baseline model that has the lowest root mean square error (RMSE). Second, we build another set of models by excluding one distance measure at a time from the baseline model. Our argument is that the resultant change in RMSE represents the predictive power of the excluded distance measure. Excluding the distance metric with the highest predictive power should lead to the largest increase in the RMSE.

5. Results

5.1. Cluster analysis

The cluster analysis using the X-means algorithm revealed four underlying clusters[4] (Table 3). Among these, we particularly focus on the clusters with the lowest (Cluster 4) and highest (Cluster 1) average count of dyadic ties. We observe that country dyads in cluster 4 have the lowest average count of ties as well as the highest average

[1] The updated psychic distance measures are available on https://sites.google.com/site/ddowresearch/home. New country pairs were added in February 2010.

[2] Exhaustive description of the differences between predictive and explanatory modeling is beyond the scope of this paper. For more see Shmueli & Koppius [35]) and Shmueli [34].

[3] This is the default option in the X-means module of RapidMiner, which is the machine learning platform we used for the analysis.

[4] As stated earlier, the variables were normalized prior to clustering. However, to obtain the average values in their original scales, the centroid tables were calculated with de-normalized values.

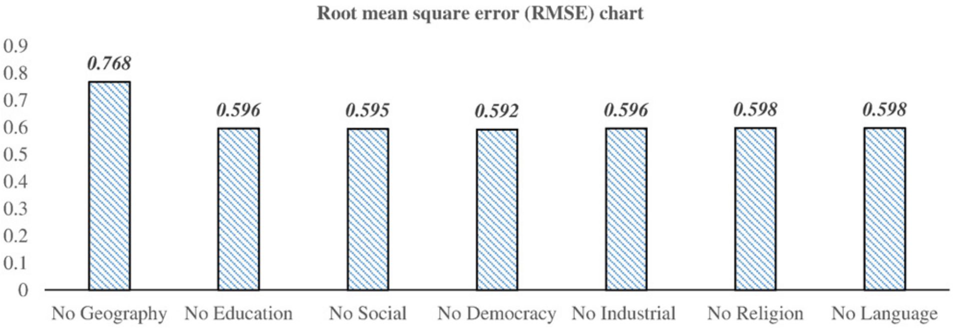

Fig. 2. Predictive power of each distance metric (regression tree model).

geographic distance. Inversely, cluster 1 includes country dyads that have the highest average count of ties and the lowest average geographic distance. Collectively, these clusters indicate a possible inverse association between geographic distance and the count of dyadic ties.

5.2. Robustness checks: cluster analysis

We test the stability of the clusters by varying the composition of the dataset in several ways. First, we examine whether most represented countries drive the cluster results. The behavior of individuals from an overly represented country could affect the overall behavioral pattern [10]. Therefore, we exclude all the observations in which, either the source or the destination country was Germany, United Kingdom, Italy, or France. We chose these countries because, during the observation period, over 35% of active traders were from these three countries. After rerunning X-means, we obtain another 4 cluster solution (Table 4). While the cluster with the highest average count of ties is no longer associated with lowest average geographic distance (Cluster 1), country dyads in cluster 4 have the highest average geographic distance as well as the lowest average count of ties. Hence, we conclude that the latent group of country dyads, characterized by highest average distance and lowest average network tie count, is robust to the inclusion of overrepresentation of countries.

Next, we incorporate two additional, platform-specific measures for each country pair. Specifically, we include the total number of investors from each country in a given dyad. We calculate these variables as follows: for each country c, we first obtain the number of distinct users that were present on the platform for the first observation week (x_{c1}). We use this number as the baseline representation of country c. Then, for each subsequent week t, we count the number of new investors who joined in week t (x_{ct}). Finally, we compute a country's cumulative presence in period $t+1$ as follows:

$$x_{ct+1} = x_{c1} + \sum_{k=2}^{t} x_{ck}$$

Table 5 shows the 4 clusters obtained after including these two new predictors. The output is qualitatively similar to the one reported in Table 4. That is, the country dyads with the lowest average count of ties are clustered with the highest average geographic distance (Table 5).

As the final robustness check, we test whether the clusters are sensitive to the exclusion of one or more observations. We generate two subsamples by excluding 30 and 50% randomly selected observations from the entire dataset. We run X-means for each subsample. For expositional brevity, we only report the results obtained using the 50% of the dataset (Table 6). However, the results are consistent across all the three subsamples. Once again, we find a 4-cluster solution similar to the one in Table 3. The country dyads with the highest (lowest) average count of ties are clustered with lowest (highest) average geographic distance.

In sum, cluster analysis reveals two robust patterns. First, we find that there exists a latent group of country dyads which has highest average geographic distance as well as the lowest average count of ties, suggesting that in settings in which users' affiliations to a particular market side do not constrain the choice of tie partners, geographic distance is still pertinent. Second, the majority of the psychic distance measures do not exhibit a similar pattern. That is, in none of the clusters, country dyads with the lowest (highest) average count of ties have the highest (lowest) average value of the psychic distance measures. The only exception is democracy dimension (part of the political distance). This lack of consistent pattern other distance measures highlights the salience of the association between the geographic distance and count of ties.

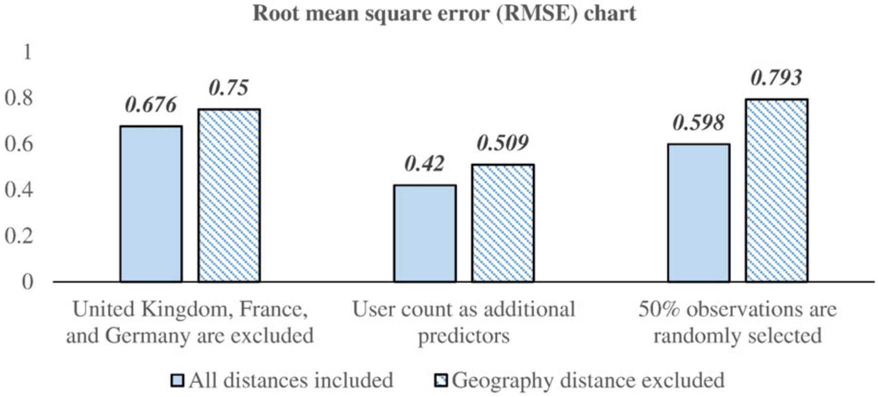

Fig. 3. Robustness checks for predictive model (regression tree). Note: for expositional brevity, we have not provided the change in RMSE after dropping the other distance measures.

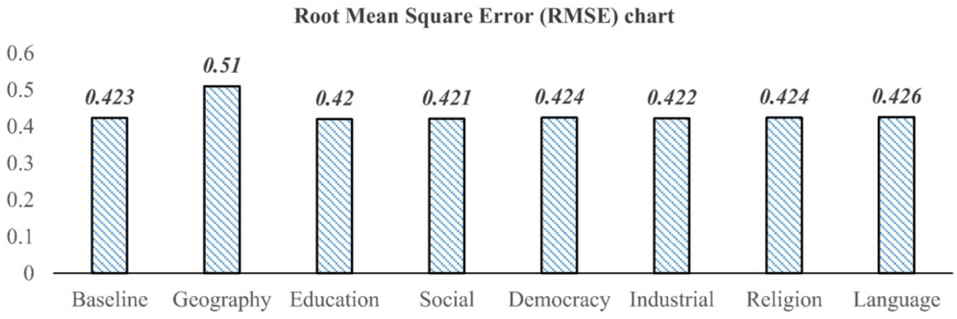

Fig. 4. Predictive Power of each distance metric (interactions between psychic and geographic distances included).

5.3. Predictive modelling: building the baseline model (step 1)

While the cluster analysis indicates that geographically most distant country dyads have the lowest average number of ties, it does not reveal the extent to which geographic distance, compared other distance measures, predicts occurrences of ties. To this end, we construct a series of the predictive model using a variety of modelling approaches. Specifically, we employ regression tree [40], pooled linear regression, k-nearest neighbor (k-NN) with five neighbors [24], and multi-layer perceptron with one intermediate layer [12]. These techniques are suitable because they allow continuous outcome variable.

As discussed in Section 4.2, we conduct the predictive modelling in two phases. First, for each technique, we use geographic and psychic distance measures to predict the count of ties between country dyads. For each technique, we run five-time five-fold cross validation [7]. In each iteration, the dataset is randomly split into five subsets. Out of these, one subset is used for testing while the remaining ones are used for training. A single iteration continues until each of five subsets has been used as a testing dataset. At the end of five such iterations, the average value of the root mean square error (RMSE) captures the extent to which the model's predictions deviate from the actual values. The exact formula for RMSE is as follows:

$$\text{RMSE} = \sqrt[2]{\frac{\sum_{I=1}^{N}\left(\hat{Y}-Y\right)^2}{N}}$$

$\hat{Y}$ and Y respectively refer to the predicted and the actual values of the outcome variable. N indicates the total number of observations. Fig. 1 compares the predictive accuracies of the techniques used. We find that among the four techniques, regression tree has the lowest RMSE (0.598). Therefore, we choose regression tree as the primary modelling approach for teasing out the predictive power of each distance measure.

5.4. Predictive modelling: comparing the predictive powers of distance measures (step 2)

To compute the predictive strength of each distance measure, we adopt the following approach: starting with the regression tree model built in step 1 as the baseline, we remove one of the distance measures. We rerun the regression tree model using five times five-fold cross validation and obtain the RMSE score. We compare this newly computed RMSE score with that of the regression tree baseline model (calculated using all the distance measures). The rationale is that the change in RMSE (i.e. RMSE of the new model in step 2 – RMSE of the baseline model in step 1) indicates the predictive power of the omitted distance measure. The distance measure with the most (least) predictive power will demonstrate the highest (lowest) increase in RMSE. We repeat this process for each distance measure (geographic distance and six measures of psychic distance), creating a total of seven regression tree models in step 2. We report the RMSE values for each model in Fig. 4. We observe that RMSE value increases by 28% after dropping geographic distance (from 0.598 in the baseline model in step 1 to 0.768 in step 2). Evidently, across all the models computed in step 2, excluding the geographic distance results in the highest change in RMSE (Fig. 2). Therefore, we conclude that among all the distance forms, the geographic distance is the strongest predictor of the count of ties between two countries.

5.5. Robustness checks: predictive modelling

We subject the predictive model results to several robustness checks. First, we conduct the checks similar to the ones reported in the cluster analysis (i.e. excluding the highly represented countries, including the count of unique users in each country as additional predictors, and randomly sampling 50% of the observations as the initial dataset). For each robustness check, we develop a baseline model using regression tree and by including all the distance measures. We assess the change in RMSE after excluding each distance measure. We find that the rise in RMSE is consistently highest when the geographic distance is excluded from the models (Fig. 3).[5] Therefore, we conclude that the predictive power of geographic distance for the count of network ties is robust to the presence of highly represented countries, the extent of each country's representation on the platform, and the configuration of the dataset used.

Also, we conduct two more robustness checks for predictive modelling. We include the interactions between geography and psychic distance measures as additional predictors. Literature suggests that different distance measures do not operate in isolation. For example, Burtch et al. [10] showed that geographic distance and cultural distance are substitutes of each other. Given the possible interactions between distance measures, it is acceptable to include the product of predictors as additional predictor variables [19]. We observe that adding the interaction terms indeed reduces the prediction error of the baseline model by 29% (RMSE = 0.423), suggesting that the interaction terms are valuable predictors. However, exclusion of geographic distance continues to cause the highest increase (20.5%) in RMSE (Fig. 4). Hence, we claim that the relative predictive strength of geographic distance does not depend on possible interactions between psychic and geographic distances.

As the final robustness check, we assess whether the predictive modelling results are susceptible to the choice of the distance measures. In the extant literature, several distance measures have been proposed. To test the robustness of our results to a different set of distance measures, we replace the psychic distance measures used so far [13] with distance measures developed by Berry et al. [5]. Drawing from the institutional theory perspective, the second set of distance measures captures eight different forms of distances between countries (Table 7).We choose this set of distance measures over others because of its coverage of countries. For instance, distance measures developed by Håkanson & Ambos [20] as well as by Brewer [8] only

[5] We don't report the changes in RMSE after dropping other distances but they are consistent with those in Fig. 2

Table 7
Institutional theory-based distance measures.
(The constituent measures are as described in Berry et al. [5], p. 1464).

Distance measure	Constituent measures	Mean	Standard deviation
Administrative	1. Colonial ties 2. Language 3. Religion 4. Legal systems	23.594	27.718
Financial	1. Financial sector development	4.631	3.958
Economic	1. Economic development 2. Macroeconomic characteristic	8.860	10.508
Political	1. Political stability 2. Democracy 3. Trade-bloc membership	2912.877	2648.592
Demographic	1. Demographic characteristics	11.301	10.499
Knowledge	1. Patents and scientific productions	7.139	9.575
Global connectedness	1. Tourism 2. Internet use	3.297	3.044
Cultural	1. Attitudes towards authority 2. Trust 3. Individuality 4. Importance of work and family	0.419	2.248

cover 25 countries while measures by Berry et al. [5] cover well over 100 countries (p. 1466).

After reconstructing the predictive models with the new set of measures, we find that while the RMSE of the new baseline model is higher than that of the baseline model developed using psychic distance measures, the geographic distance remains the strongest predictor of network ties. We observe that the gain in RMSE from the baseline model is highest (approximately 18%) when the geographic distance is excluded (Fig. 5). The change in RMSE when other distance measures are dropped is lower than that when geographic distance is omitted. Therefore, we conclude that geographic distance remains a superior predictor of network ties regardless of the distance measures used.

6. Conclusion

We began this study by identifying two gaps in the extant literature on distance effects in online contexts: predominance of geographic distance as a measure of distance & the resultant exclusion of other distance measures, and lack of study of online contexts that do not have a two-sided market structure. Our research objective is to assess the extent to which geographic distance, compared to alternative distance measures, predicts dyadic ties in an online context that does not follow a two-sided market structure. We find that not only is the higher geographic distance negatively associated with fewer occurrences of ties but also, compared to a range of other distance measures, geographic distance is the strongest predictor of the count of ties.

Apart from being the first comparative study of the predictive power of geographic and psychic distance measures in online context, our study makes two more contributions. First, the cluster analysis reveals that the greater geographic distance is associated with fewer occurrences of ties. However, geographically nearer countries are not consistently associated with the highest count of ties. A possible explanation is that greater geographic distance may be a sufficient to have fewer ties but lower geographic distance may not suffice to have more ties. That is, while geographic distance may have an overarching association with the count of ties, lack of geographic distance may make other distance measures pertinent. Future studies could assess whether this pattern is causal in nature.

Second, the predictive superiority of geographic distance suggests that at an aggregate level, individuals more readily perceive others to be distant/proximal on a geographic distance dimension than on other distance dimensions. This assertion seems somewhat contrary to the explanatory model which suggest that influence of geographic distance on the occurrence of ties is either substitutable with other distance measures [10] or absent altogether [32]. Therefore, additional work is required to compare the predictive and explanatory models of distance effects in online settings. However, our finding is generally consistent with the extant literature because we find evidence contrary to the "death of distance" hypothesis.

One must view our results in light of certain limitations. First, our study does not claim a causal link between the count of ties and geographic distance. This restriction exists given that predictive modelling

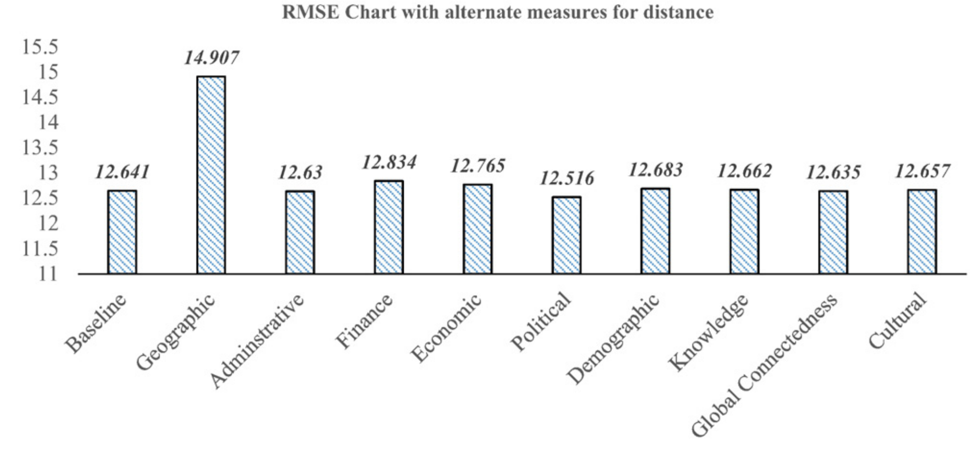

Fig. 5. Predictive power of each distance metric (alternate distance measures).

approaches are more concerned with making accurate predictions than with discussing the underlying explanatory mechanisms [35]. Second, the predictive model constructed by aggregating the data to country dyads may not automatically extrapolate to dyads between individuals. At an individual level, additional factors that are not readily available at country-level may predict the occurrence of ties. Therefore, further research is required to test whether and to what extent is geographic retains its predictive power at a more granular level. Finally, ties in our context are not costless. Therefore, users may be more sensitive to the notion of distance. Future studies can replicate the study in settings in which forming a new tie is almost costless (e.g. Facebook) to assess the generalizability of the patterns observed here.

References

[1] A. Abbasi, C. Albrecht, A. Vance, J. Hansen, MetaFraud: a meta-learning framework for detecting financial frauds, MIS Q. 36 (4) (2012) 1293–1327.

[2] A. Abbasi, Z. Zhang, D. Zimbra, H. Chen, J. Nunamaker, Detecting fake websites: the contribution of statistical learning theory, MIS Q. 34 (3) (2010) 435–461.

[3] A. Agarwal, C. Catalini, A. Goldfarb, The Geography of Crowdfunding(Retrieved from The National Bureau of Economic Research) http://www.nber.org/papers/w16820.pdf 2, 2011.

[4] R. Bapna, P. Goes, A. Gupta, Y. Jin, User heterogeneity and its impact on electronic auction market design: an empirical exploration, MIS Q. 28 (1) (2004) 21–43.

[5] H. Berry, M. Guillén, N. Zhou, An institutional approach to cross-national distance, J. Int. Bus. Stud. 41 (9) (2010) 1460–1480.

[6] B. Blum, A. Goldfarb, Does the internet defy the law of gravity? J. Int. Econ. 70 (2) (2006) 384–405.

[7] M. Bogaert, M. Ballings, D. Van den Poel, The added value of Facebook friends data in event attendance prediction, Decis. Support. Syst. 82 (2016) 26–34.

[8] P. Brewer, Operationalizing psychic distance: a revised approach, J. Int. Mark. 15 (1) (2007) 44–66.

[9] V. Burbano, Social responsibility messages and worker wage requirements: field experimental evidence from online labor marketplaces, Organ. Sci. 27 (4) (2016) 1010–1028.

[10] G. Burtch, A. Ghose, S. Wattal, Cultural differences and geography as determinants of online prosocial lending, MIS Q. 38 (3) (2014) 773–794.

[11] F. Cairncross, The Death of Distance: How the Communications Revolution Is Changing Our Lives—Distance Isn't What It Used to Be, Harvard Business School Press, Boston MA, 2001.

[12] P. Cortez, A. Cerdeira, F. Almeida, T. Matos, J. Reis, Modeling wine preferences by data mining from physicochemical properties, Decis. Support. Syst. 47 (4) (2009) 547–553.

[13] D. Dow, Centre for the Practice of International Trade-Psychic Distance(Retrieved from Melbourne Business School) https://sites.google.com/site/ddowresearch/home 2 4, 2010.

[14] D. Dow, A. Karunaratna, Developing a multidimensional instrument to measure psychic distance stimuli, J. Int. Bus. Stud. 37 (5) (2006) 578–602.

[15] P. Ellis, Does psychic distance moderate the market size-entry sequence relationship? J. Int. Bus. Stud. 39 (3) (2008) 351–369.

[16] J. Evans, F. Mavondo, Psychic distance and organizational performance: an empirical examination of International retailing operations, J. Int. Bus. Stud. 33 (3) (2002) 515–532.

[17] R. Finberg, 10 Questions About Saxo Bank's New 'Social' TradingFloor.com, What to Expect(Retrieved 01 23, 2015, from Finance Magnets) http://www.financemagnates.com/forex/brokers/10-questions-about-saxo-banks-new-social-tradingfloor-com-what-to-expect/ 1 31, 2014.

[18] D. Gefen, E. Carmel, Is the world really flat? A look at offshoring in an online programming marketplace, MIS Q. 32 (2) (2008) 367–384.

[19] I. Guyon, A. Elisseeff, An introduction to variable and feature selection, J. Mach. Learn. Res. 3 (2003) 1157–1182.

[20] L. Håkanson, B. Ambos, The antecedents of psychic distance, J. Int. Manag. 16 (3) (2010) 195–210.

[21] A. Hortaçsu, F. Martínez-Jerez, J. Douglas, The geography of trade in online transactions: evidence from eBay and MercadoLibre, Am. Econ. J. Microecon. 1 (1) (2009) 53–74.

[22] L. Jeppesen, K. Lakhani, Marginality and problem-solving effectiveness in broadcast search, Organ. Sci. 21 (5) (2010) 1016–1033.

[23] J. Johanson, J. Vahlne, The Uppsala internationalization process model revisited: from liability of foreignness to liability of outsidership, J. Int. Bus. Stud. 40 (9) (2009) 1411–1431.

[24] Y. Lee, C. Wei, T. Cheng, C. Yang, Nearest-neighbor-based approach to time-series classification, Decis. Support. Syst. 53 (1) (2012) 207–217.

[25] B. Lengyel, A. Varga, B. Ságvári, Á. Jakobi, J. Kertész, Geographies of an online social network, PLoS One 10 (9) (2015) (Retrieved from http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0137248).

[26] M. Li, A. Kankanhalli, S. Kim, Which ideas are more likely to be implemented in online user innovation communities? An empirical analysis, Decis. Support. Syst. 84 (2016) 28–40.

[27] M. Lin, S. Viswanathan, Home bias in online investments: an empirical study of an online crowdfunding market, Manag. Sci. 62 (5) (2016) 1393–1414.

[28] T. Mayer, S. Zignago, Notes on CEPII's Distances Measures: The GeoDist Database(Retrieved from Social Science Research Network) http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1994531 12 1, 2011.

[29] G. Meyer, G. Adomavicius, P. Johnson, M. Elidrisi, W. Rush, J. Sperl-Hillen, P. O'Connor, A machine learning approach to improving dynamic decision making, Inf. Syst. Res. 25 (2) (2014) 239–263.

[30] A. Ojala, Geographic, cultural, and psychic distance to foreign markets in the context of small and new ventures, Int. Bus. Rev. 24 (5) (2015) 825–835.

[31] D. Pelleg, A. Moore, X-means: extending K-means with efficient estimation of the number of clusters, Proceedings of the Seventeenth International Conference on Machine Learning, ACM, San Francisco 2000, pp. 727–734.

[32] O. Posegga, M. Zylka, K. Fischbach, Collective dynamics of crowdfunding networks, 48th Hawaii International Conference on System Sciences, IEEE, Kauai 2015, pp. 3258–3267 (Retrieved from http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7070209&isnumber=7069647).

[33] A. Rai, X. Tang, P. Brown, M. Keil, Assimilation patterns in the use of electronic procurement innovations: a cluster analysis, Inf. Manag. 43 (3) (2006) 336–349.

[34] G. Shmueli, To explain or to predict, Stat. Sci. 25 (3) (2010) 289–310.

[35] G. Shmueli, O. Koppius, Predictive analytics in information systems research, MIS Q. 35 (3) (2011) 553–572.

[36] L. Wallace, M. Keil, A. Rai, Understanding software project risk: a cluster analysis, Inf. Manag. 42 (1) (2004) 115–125.

[37] V. Wohlgemuth, E. Berger, M. Wenzel, More than just financial performance: trusting investors in social trading, J. Bus. Res. 69 (11) (2016) 4970–4974.

[38] H. Yildiz, C. Fey, Are the extent and effect of psychic distance perceptions symmetrical in cross-border M&As? Evidence from a two-country study, J. Int. Bus. Stud. 47 (7) (2016) 830–857.

[39] A. Zaheer, E. Hernandez, The geographic scope of the MNC and its alliance portfolio: resolving the paradox of distance, Glob. Strateg. J. 1 (1–2) (2011) 109–126.

[40] Y. Zhao, Y. Zhang, Comparison of decision tree methods for finding active objects, Adv. Space Res. 41 (12) (2008) 1955–1959.

[41] Y. Takhteyev, A. Gruzd, B. Wellman, Geography of Twitter networks, Soc. Networks 34 (1) (2012) 73–81.

Swanand J. Deodhar is a PhD candidate in business administration at the Carlson School of Management, University of Minnesota. His areas of research interest are crowdsourcing, evolution of online network ties, and collective intelligence. He is a student member of Association of Information Systems, INFORMS, and the Academy of Management.

Mani Subramani is an associate professor in the Department of Information and Decision Sciences at the Carlson School of Management, University of Minnesota. His research interests include the management of relationships between organizational groups and between organizations, knowledge management, and electronic commerce. His paper titled the "Dot Com Effect: The Impact of E-Commerce Announcements on the Market Value of Firms" won the Best Paper award at the 20th International Conference on Information Systems in December 1999. He has published in many journals, including MIS Quarterly, Information Systems Research, Academy of Management Journal, and Journal of Management Information Systems.

Aks Zaheer received his PhD in strategic management from the Massachusetts Institute of Technology and his Master's in Business from the Indian Institute of Management in Ahmedabad. His current research examines the antecedents and consequences of trust in organizations and in interfirm exchange, strategic alliances, mergers and acquisitions, and the dynamics of social structure in organizations, among others. He has published in many journals, including Administrative Science Quarterly, Organization Science, Strategic Management Journal, Academy of Management Review, and Academy of Management Journal. He is an elected Fellow of the Strategic Management Society. He was named the Outstanding Core Teacher of the Year for the Full-Time MBA Program in 1995, 2005, 2009, 2010, and 2011; and received the Curtis Cup Outstanding Teacher Award in 2006 and 2015 for Executive MBA teaching, Excellence in Teaching Awards for 2004, 2009, 2012 and 2014, Outstanding Research Award in 2014, commended for teaching excellence in Business Week's guide to the 50 Top Business Schools in 1997 and named one of the World's 50 Best B-School Professors by Poets and Quants in 2012.