Determinants of Social Connectedness in India and Implications for Social Learning*

Vasu Chaudhary

Abstract

This paper utilizes the de-identified and aggregated Facebook friendship network across Indian geography to explore which factors determine the patterns of social connectedness. I find that geographical distance and state boundaries can play a significant role in influencing network patterns, and the elasticity increases as we look at regions located closer to each other. I also showcase the existence of homophily along the dimensions of literacy, language, caste, and religion. Through a set of simulations (DeGroot learning), I further look at the implications of such network structure on the speed of learning.

**Corresponding address: chaudhary.vasu.27@gmail.com. Acknowledgments: I am grateful to my peers for helpful discussions. All errors are my own.

1 Introduction

Humans are inherently social beings. From the beginning of humankind, we have functioned as part of a society where people are engaged in continuous interaction in one dimension or the other. Social networks play an essential role in every aspect of our lives. This becomes even more crucial in developing economies where access to formal institutions is scarce and informal networks influence information flows, political polarization, business decisions, trade, migration, and a myriad of other social and economic activities. In addition, the data revolution that the developing economies are going through accentuates the role that social networks end up playing.

This paper analyzes the de-identified and aggregated Facebook friendship network across the Indian geography. I aim to show how spatial and socio-economic factors can explain the variation in social connectedness over the platform. Empirical research on such friendship networks - their determinants and effects - has been done in the past by researchers (Bailey et al. 2018; Bailey, Farrell, et al. 2020; Bailey, Gupta, et al. 2020; Bailey, Kuchler, et al. 2020); however, the focus has always been limited to developed economies of the west. There have been two primary reasons for the same. First, the developing world experienced digital penetration at a rate much slower than that of the U.S. and Europe. Nevertheless, with the rapid growth in smartphone and internet access, this context is changing, and now people in the developing world are moving beyond the pre-existing barriers in how and with whom they wish to connect. Second, historically it has been challenging to get access to microdata on how people connect at such a large scale. This barrier is finally being brought down with data becoming more accessible (albeit at an aggregated level) to the broader research community.

India has been experiencing rapid growth in social media usage over the years, with the active users jumping from 135 million in 2015 to almost 350 million in 2020, with further rise predicted in the future (Figure 5 panel A). Given India has a large proportion of the young population, it is no surprise that most users belong to age groups 18-24 and 25-34

(Figure 5 panel B). The impact that social media usage can have in the Indian context (Facebook in particular) has been explored in the past along dimensions such as the impact on self-empowerment (Kumar 2014), addiction and loneliness amongst the youth (Shettar et al. 2017), disaster management (Bhuvana and Aram 2019), and electoral outcomes (Barclay et al. 2015). This, in conjunction with the growing reach in the sub-continent, calls for a need to explore what factors determine the structure of such networks. Bailey et al. (2018) were the first to explore the determinants of social connectedness in the U.S. and later similar analysis for Europe in Bailey, Kuchler, et al. (2020) . Their results highlighted the role that geographic distance and state boundaries can play in influencing social connectedness and how homophily (tendency for people to link with those similar to themselves) exists along the dimensions of education, age, language, and religion. I extend their framework to the Indian context, and as one would expect apriori, we see a strong influence of spatial factors here. The results also showcase the impact of linguistic, caste, and religious differences on social media. Further, I take a brief detour into the realm of social learning and explore some implications of the structure of these social networks across Indian districts.

This paper is organized as follows. Section 2 describes the Facebook network data I utilize and other covariates from the India Census. Section 3 presents the results for the determinants of social connectedness in India. Section 4 describes the implications of network structure on social learning. Section 5 concludes.

2 Data and Methodology

I utilize the anonymized and aggregated snapshot of active Facebook users (*available on Facebook data for good*) across Indian districts – an administrative level similar to counties in the U.S. The data is provided in the form of a Social Connectedness Index (henceforth SCI), which measures the intensity of friendship links between each district pair, as of August 2020. To state it precisely, the SCI between any two locations i and j is given by the following –

$$\text{Social Connectedness Index}_{ij} = \frac{\text{FB Connections}_{ij}}{\text{FB Users}_i * \text{FB Users}_j} \tag{1}$$

This formulation ensures that the index takes into account that locations with more users will be more likely to be highly connected. The SCI_{ij} measures the relative probability of a Facebook link between a Facebook user in location i and location j. Hence, we can only interpret the relative magnitudes – twice the measure implies that a Facebook user in location i is twice as likely to be connected with a Facebook user in location j. Also, Facebook requires that one user send a request and the other accept it (unlike Twitter), hence the network is undirected.

To give an overview of how one may represent the SCI, Figure 1 shows the distribution of friendship networks for Pune district, Maharashtra. In a fashion similar to Bailey, Kuchler, et al. (2020) , I represent the intensity of cross-district links scaled by the 20th percentile of overall SCI across all district pairs. At first glance, we can see geographical clustering in the network, with the intensity of connections falling within distance. Results in the next section further confirm this and highlight how other demographic factors correlate with the SCI.

Using the Indian census of 2011, I extracted socio-economic and demographic data at the district level and calculate variables for each district pair. These demographic indicators measure differences in literacy rates across districts, the difference in urbanization levels, the difference in religious composition, the difference in caste composition, and whether the district pair shares the same language spoken by a majority. These demographic characteristics can be exploited to explain why groups (districts in our setup) are more likely to be linked with other groups sharing similar traits. Extensive research on homophily in social networks has been done in the past (Currarini, Jackson, and Pin 2009; Currarini, Matheson, and Vega-Redondo 2016; McPherson, Smith-Lovin, and Cook 2001). The prevalence of homophily can be a crucial factor in explaining opinion dynamics in a society (Golub and Jackson 2012).

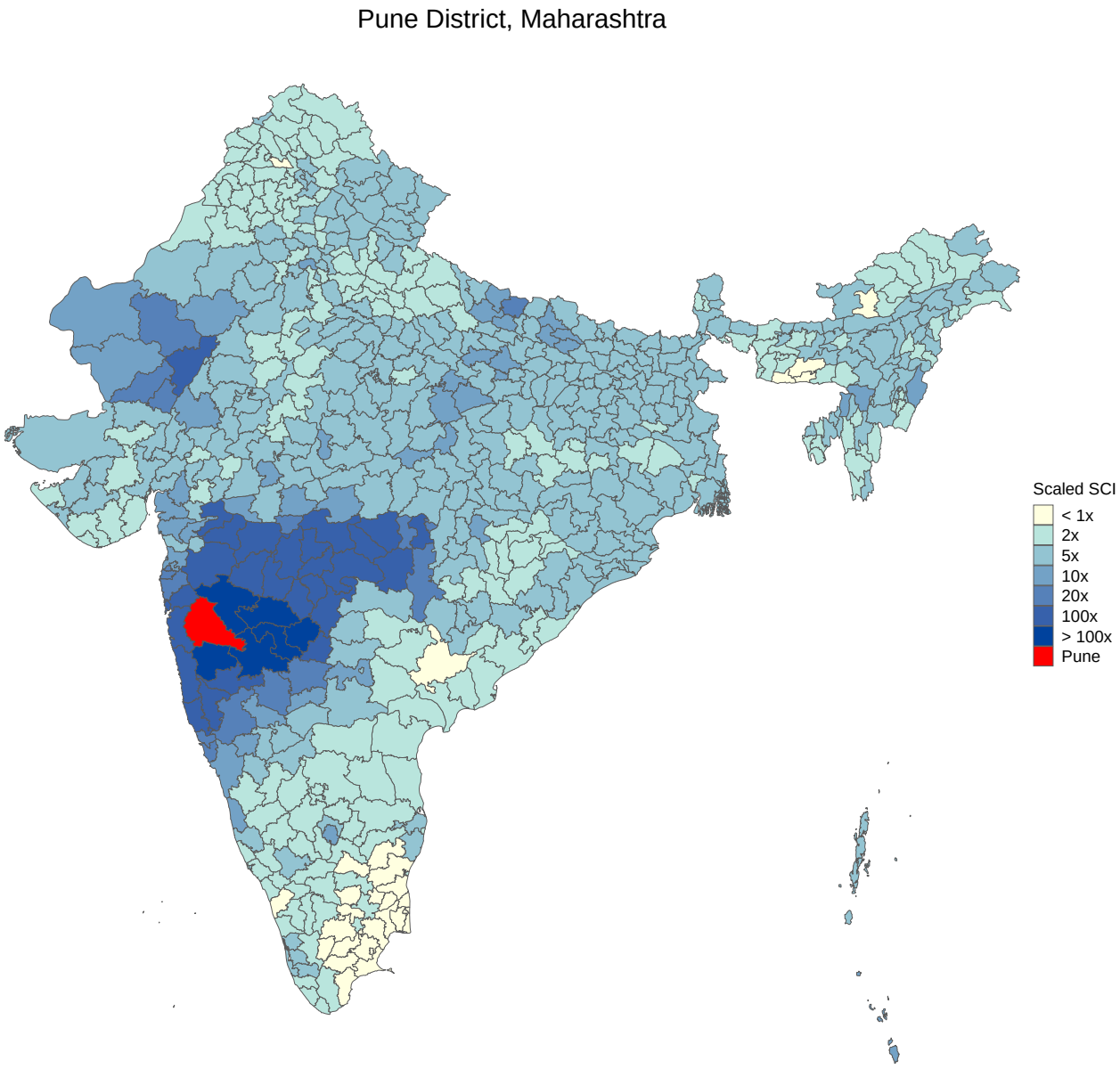

Figure 1: Mapping Social Connectedness Index for Indian Districts. The plot shows relative probability of connections between the Pune district in red with all the other districts.

3 Results

To determine which factors are associated with social connectedness across Indian districts, I begin by estimating the following regression:

$$log(SCI_{ij}) = \alpha + \beta log(d_{ij}) + \gamma X_{ij} + \delta_i + \delta_j + \epsilon_{ij} \tag{2}$$

Where i and j refer to the district pairs across Indian administrative regions, SCI_{ij} is the social connectedness index, d_{ij} measures the distance between each district pair (I use the centroid for each polygon to calculate these distances). The vector of covariates X_{ij} contains measures of similarity as discussed in the previous section: caste composition, literacy rate, primary language, religious composition, urbanization rates. Also, I add fixed effects for each district pair.

Table 1 includes the results of this estimation. Column 1 includes only distance (in meters), and we can see as expected that users in districts that are farther apart in space have a relatively lower probability of being connected compared to those located closer. The magnitude of elasticity indicates that a 10% increase in distance is associated with a 14.2% decrease in the social connectedness index. The high R^2 shows that distance indeed explains a large proportion of variation in cross-district links. If we were to compare this with the results from a similar analysis in the western nations (Bailey, Kuchler, et al. 2020), the distance explains more variation in the Indian context than Europe (36%). In column 2, I also add a control for whether the two districts belong to the same state. We see that being in the same state leads to connectedness being 3.5 times higher and is statistically significant. This magnitude falls marginally when we account for additional demographic factors in column 3. We can see that having the same majority language corresponds to a 40% increase in connectedness across districts. Also, a wedge in literacy levels correlates to a fall

Table 1: Determinants of Social Connectedness in India

	Dependent Variable: log(SCI)					
	(1)	(2)	(3)	(4)	(5)	(6)
log(Distance)	-1.421***	-1.212***	-1.105***	-1.605***	-0.898***	-1.240***
	(0.075)	(0.085)	(0.059)	(0.069)	(0.080)	(0.059)
Same State		1.312***	1.158***	0.897***	1.174***	
		(0.224)	(0.183)	(0.117)	(0.296)	
Language			0.373***	0.660***	0.347***	0.603***
			(0.092)	(0.130)	(0.083)	(0.117)
Δ literacy			-0.702***	-0.874**	-0.565***	-0.795***
			(0.230)	(0.336)	(0.207)	(0.223)
Δ Median age			0.026***	0.021**	0.017**	0.024**
			(0.008)	(0.009)	(0.007)	(0.010)
Δ Urban share			0.277***	0.101	0.164***	0.302***
			(0.075)	(0.136)	(0.056)	(0.074)
Δ SC share			-0.537**	-0.552**	-0.387	-0.561**
			(0.240)	(0.262)	(0.255)	(0.244)
Δ ST share			-0.487**	-1.043***	-0.337**	-0.515***
			(0.178)	(0.308)	(0.129)	(0.144)
Δ UC share			-0.014	0.510**	0.062	-0.006
			(0.102)	(0.237)	(0.055)	(0.065)
Δ Hindu share			-0.774***	-0.169	-0.603***	-0.739***
			(0.198)	(0.157)	(0.169)	(0.179)
Δ Muslim share			-0.033	-0.589***	-0.165	-0.314
			(0.180)	(0.213)	(0.139)	(0.201)
Cut-off				< 350 km	> 350 km	
District FE	x	x	x	x	x	x
State FE						x
N	420538	420538	371490	38890	332600	371490
R2	0.722	0.761	0.791	0.876	0.669	0.763

* p < 0.1, ** p < 0.05, *** p < 0.01

[1] Distance is measured between all district centroid pairs. Social Connectedness Index between districts i and j looks at the relative probability of friendship links between them. Standard errors provided in the parantheses are double clustered at the level of states for the district pair.SC = Scheduled Caste, ST = Scheduled Tribe, UC = Upper Caste.

in connectedness. A percentage point increase in the difference in literacy rate for the district pair leads to a 50% fall in connectedness. If we look at the rates of urbanization, we see an exciting outcome. An increase in the urbanization gap is associated with a significant rise in levels of social connectedness. One explanation for this result could be the high migration rates from rural to urban India (Dubey, Palmer-Jones, and Sen 2006; Mitra and Murayama 2009).

Next, we look at the impact caste composition can have. India has a deep-rooted history of segregation along caste lines, and it can be an essential factor in explaining homophily (Hoff and Pandey 2004; Sidhwani 2015). We see that deviation in Scheduled Caste and Scheduled Tribe composition is significantly associated with a fall in connectedness. A percentage point increase in the gap between the Scheduled Caste population across a district pair implies a 45% decrease in connectedness, and a percentage point increase in a similar gap for Scheduled Tribe composition is associated with a 38% fall in connectedness. The coefficient for upper castes is insignificant as well as very low in magnitude. Like caste composition, we also try to evaluate religion's role in explaining variation in connectedness. We can see in column 3 that deviation in the Hindu population has a significant impact on social connectedness, with a percentage point increase in Hindu population share difference between two districts leading to a 54% fall in connectedness. These religious and caste-based underpinnings for homophily are crucial for the Indian context, given how the sub-continent has a history of communal tension. This, in conjunction with widespread misinformation diffusing through social media, can have significant consequences (Farooq 2017; Phartiyal S 2018).

In the next two columns, we split the sample into those district pairs located within 350 km of each other and those located farther than 350 km. The reason for doing so is to see how results vary when we move further apart in the geographical space. The results from columns 4 and 5 show that the impact of distance on social connectedness is much stronger for districts that are located closer to each other - the elasticity almost doubles (from -0.8 to -1.6) when we restrict the analysis to pairs within 350 km. Further, the impact of belonging to the same state is significant and positive across both the samples, reaffirming that even when users are situated in proximity, the state boundaries play a vital role. The literacy gap coefficient remains significant across both the sub-samples - the impact becomes stronger for those districts located in proximity to each other. The coefficient for urbanization share becomes insignificant when we restrict the sample to districts located close to each other. This could be an outcome of urban and rural regions located in geographical clusters; hence

we should expect less variation for such district pairs. An interesting outcome is that we see no more significant impact of the gap in the share of the Hindu population when we look at regions located close to each other. This could be an outcome of both less variation in religious population in geographical clusters, and other regional identities taking precedence. Simultaneously, in this specification, we also see the share of the Muslim population becoming significant, and a percentage point increase in the gap between the share of the Muslim population corresponds to a 44% fall in connectedness between district users. This could be indicative of minority identity becoming stronger for regions situated close to each other. This effect can be seen even for the Scheduled Tribe population share. Here, we can see a jump in effect size: from a percentage point increase in Scheduled Tribe population share gap leading to 34% decline in connectivity earlier, to now a 64% fall.

Finally, in column 6, we drop the dummy for regions belonging to the same state and add in fixed effects for each state pair. The direction and the significance of the results remain valid even under this specification, with minimal changes in magnitudes. One crucial point to keep in mind while studying these effects is that a significant proportion of variation in connectedness can be explained by the variation in geographic distance itself, and even though the effects of demographic characteristics seem significant (with relevant magnitudes), they add only a little to the overall variation. Also, these results do not imply causality but highlight the existing patterns and correlations.

The strength of weak ties was highlighted in Granovetter (1983). Having diverse connections is associated with more opportunities and greater integration. Bailey et al. (2018) showed how the dispersion of friendship links across U.S. counties is correlated with social and economic outcomes at the county level. While such an analysis can be done for Indian districts, the lack of detailed data restricts us from doing so. However, as a quick glance, Figure 2 highlights the difference in the spread of friendship networks for two districts in Karnataka. In Panel A, we have the Bangalore district, one with the state's highest income levels and has ties

relatively spread throughout Indian geography. On the other hand, in Panel B, we have the Koppal District. It is one of the poorest districts in the region, and we can see that its ties are more strongly clustered and restricted to the local geography only.

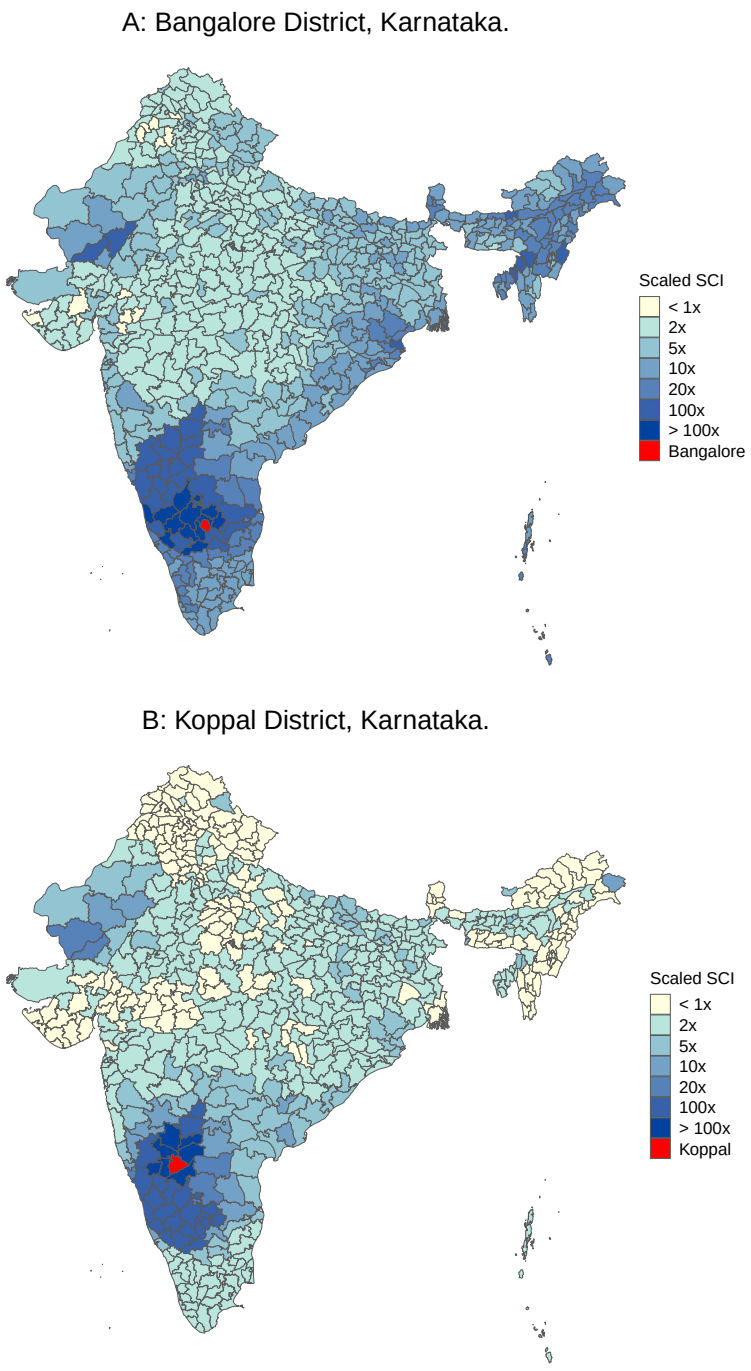

Figure 2: Variation in Social Connectedness and Income Levels

4 Social Learning

The structure of a network can have important implications for opinion dynamics in society. It can impact a wide range of outcomes: whether the group of agents converge in their beliefs, the pace at which such convergence will take place; do these dynamics vary across societies and cultures; whether a subset of agents manipulate such convergence, and what remedies can be offered. Given how much information exchange takes place over social media, understanding network structure's implications becomes crucial. Extensive literature already exists which looks at such dynamics (Chatterjee and Seneta 1977; DeGroot 1974; French Jr 1956; Hegselmann, Krause, and others 2002; Lorenz 2007).

I utilize the DeGroot repeated learning model to explore the implications of the structure of the Facebook network across Indian districts. This exercise is purely for exposition purposes and tries to showcase how strong geographical clustering and correlated beliefs can impact society's pace of convergence.

We start with a set of n agents (each district is equivalent to one node). Each agent i starts with an initial opinion estimate, described by a real number x_i. All agents then update their beliefs simultaneously at discrete times $t = 1, 2, 3\ldots$. So, at each time period t, we have a vector of opinions $x(t) = (x_1(t), x_2(t), \ldots.., x_n(t))$, and agent i's opinions can be represented as-

$$x_i(t) = \sum T_{ij} * x_j(t-1) \tag{3}$$

Here, T is a non-negative row-stochastic matrix ($\sum T_{ij} = 1$ for each i), with T_{ij} representing the weight agent i places on agent j while updating their opinion. In our case, the weights are calculated from the Social Connectedness Index for each district pair. Note that all the districts are connected with a distinct edge in this dataset, i.e., we have a complete network

with 644 nodes. Since our network is strongly connected, we already know that a consensus should arise using standard results from Markov Chain theory. The objective here is to run simulations to show how homophily we described in the previous section impacts speed of learning (Golub and Jackson 2012).

First, we start with a scenario where all districts have initial opinions drawn from a uniform distribution on $[0, 1]$. This can represent opinion/belief on a wide array of real-world scenarios such as religiosity, political spectrum, policy views. Using the process of opinion dynamics given in (3), we get the results in panel A of Figure 3. Since the network is complete and initial opinions were uniformly distributed, we see quick convergence to the average opinion. However, as we saw in our results from Table 1, there is a strong tendency of geographical concentration in social connectedness. In panel B of Figure 3, we rerun the same model given in (3), but now the initial opinions are clustered at the State level, i.e., district pairs belonging to the same state have opinions closer to each other in space. The reasoning for this is that India has a history of experiencing considerable differences along cultural, linguistic, social, economic, and political dimensions across states (Dharmalingam and Morgan 2004). This clustering of initial opinions, in conjunction with state boundaries (as we have seen in our results) also acting as essential determinants in social connectedness, gives us a slower convergence rate, as can be seen in Panel B of Figure 3.

In the simulations above, the matrix T was constant, and the agents (districts) had enough confidence (each neighbor's opinion was taken into account while updating, regardless of the magnitude of differences). Next, to highlight how lack of confidence can impact convergence, I briefly explore a non-linear class of opinion dynamics – models with bounded confidence (Hegselmann et al. 2002; Krause 2000). Here, each agent (district) will put a positive weight on only those neighbors who are *not too far away* (ϵ distance) from their own opinion. This amounts to the weight matrix dynamically updating based on the opinion vector in the previous time period. So now, the opinion dynamics are governed by the following process –

$$x(t) = T(x(t-1)) * x(t-1) \tag{4}$$

$$\text{where } T_{ij}(x(t-1)) = \begin{cases} \frac{SCI_{ij}}{\sum SCI_{ik}} & \text{if } |x_i(t-1) - x_j(t-1)| < \epsilon \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

The initial opinion vector is the same as Figure 3 panel B, with values clustered at the State level. Results are presented for three confidence bounds (*low, medium, and high*) in each of the three panels of Figure 4. As expected, higher confidence bound (Panel A) leads to faster convergence and is headed towards consensus. On the other hand, lower confidence bounds (Panel C) leads to a slow convergence rate and can result in the emergence of opinion clusters even after convergence moves towards stabilization. The intention here is to show how a lack of confidence in opinions of agents significantly different from one's own, can lead to rapid fall in the speed of learning, as well as, impede consensus altogether.

This exercise gives a brief picture of how the structure of connectedness in online networks can have implications for social learning. Further exploration with more relevant microdata combined with high-dimensional experiments in big-data environments can provide much-needed insights into theoretical models' predictive powers.

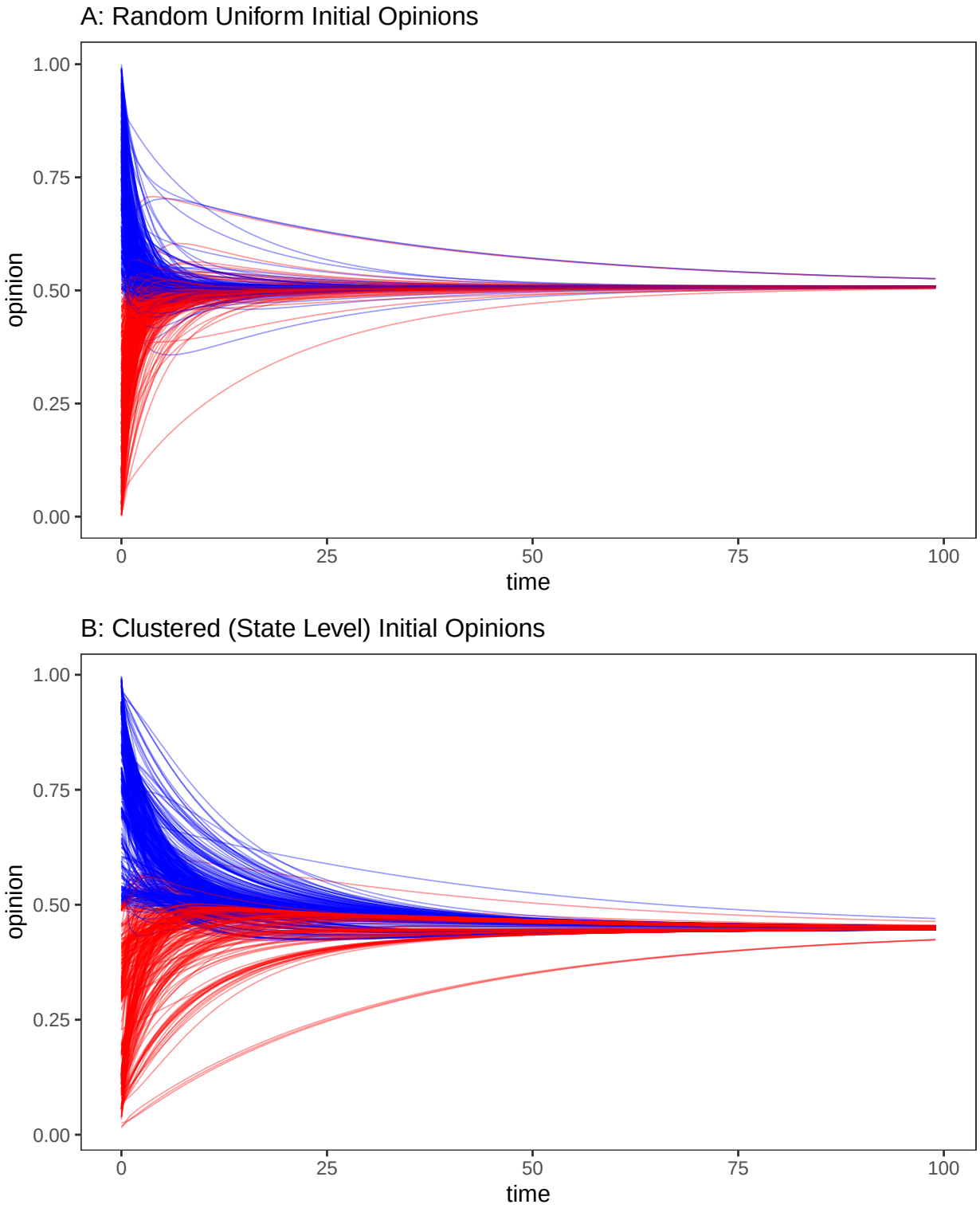

Figure 3: DeGroot repeated learning models.

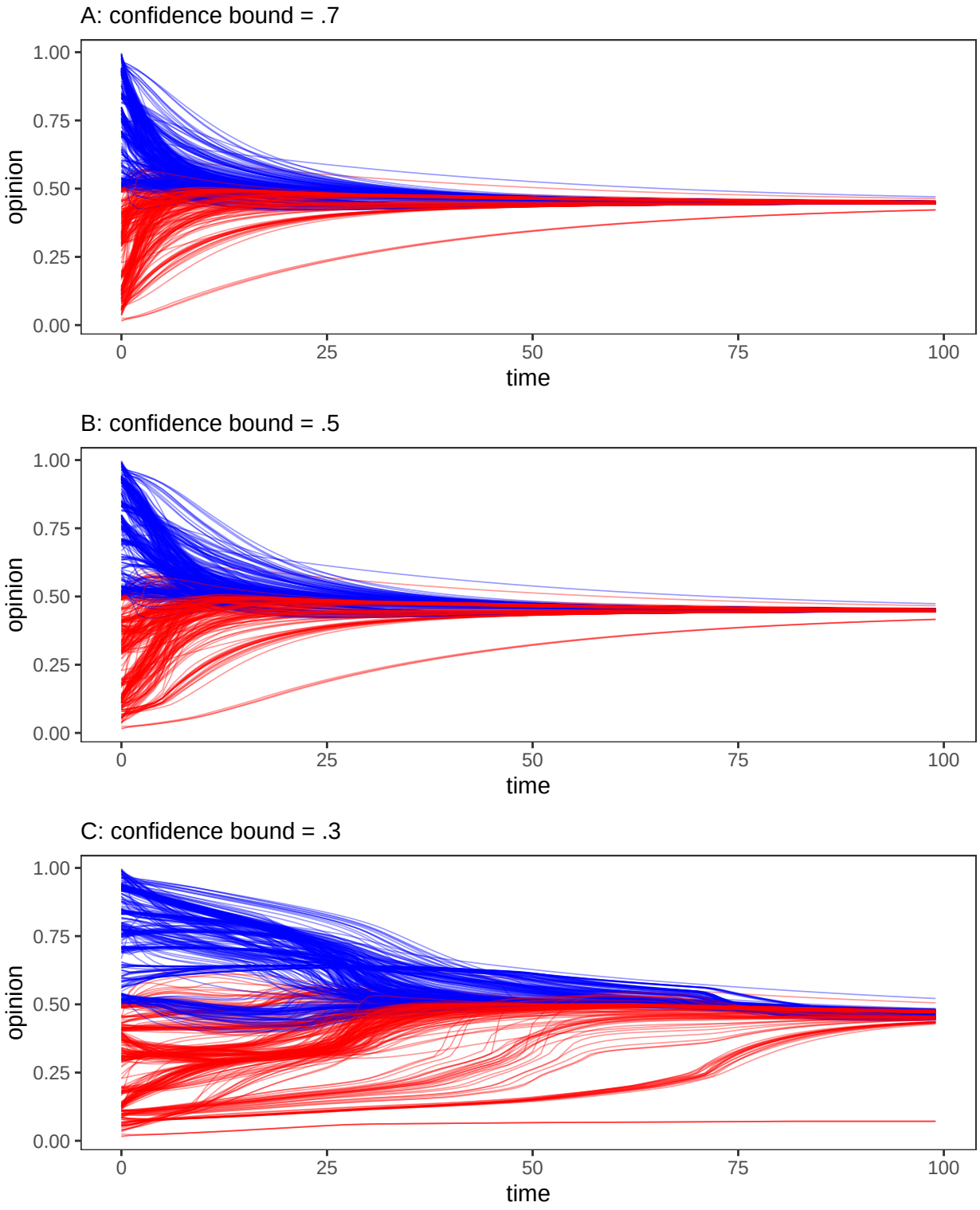

Figure 4: Opinion dynamics with confidence bounds

5 Conclusion

As the world develops further and technology starts penetrating even the darkest corners of the globe, the geographical barriers that humans have historically confronted start fading away. This paper was an attempt to unravel the determinants of social connectedness in this context for India. The results indicate that geography still plays a crucial role in influencing the structure of connections in the online space. Distance between users ends up explaining a major proportion of the variation in network structure. This is further strengthened by the discontinuity we see at the state boundaries, implying that state-specific characteristics can influence network formation. Also, I show how the intensity of social connectedness is correlated with divergence in demographic and socio-economic characteristics. These patterns of homophily in the network can have crucial effects on the speed of learning in society. Economics has been at the forefront of expanding its frontier, with technological advances continuing our reach to more complex systems. As part of future research, we need to explore further the exact dynamics of how information flows through the veins of such networks, and what interventions can help nudge society towards a better equilibrium.

6 References

Bailey, Michael, Rachel Cao, Theresa Kuchler, Johannes Stroebel, and Arlene Wong. 2018. "Social Connectedness: Measurement, Determinants, and Effects." *Journal of Economic Perspectives* 32(3):259–80.

Bailey, Michael, Patrick Farrell, Theresa Kuchler, and Johannes Stroebel. 2020. "Social Connectedness in Urban Areas." *Journal of Urban Economics* 103264.

Bailey, Michael, Abhinav Gupta, Sebastian Hillenbrand, Theresa Kuchler, Robert J. Richmond, and Johannes Stroebel. 2020. *International Trade and Social Connectedness*. National Bureau of Economic Research.

Bailey, Michael, Theresa Kuchler, Dominic Russel, Johannes Stroebel, and others. 2020. "Social Connectedness in Europe."

Barclay, Francis P., C. Pichandy, Anusha Venkat, and Sreedevi Sudhakaran. 2015. "India 2014: Facebook 'Like'as a Predictor of Election Outcomes." *Asian Journal of Political Science* 23(2):134–60.

Bhuvana, N. and I. Arul Aram. 2019. "Facebook and Whatsapp as Disaster Management Tools During the Chennai (India) Floods of 2015." *International Journal of Disaster Risk Reduction* 39:101135.

Chatterjee, Samprit and Eugene Seneta. 1977. "Towards Consensus: Some Convergence Theorems on Repeated Averaging." *Journal of Applied Probability* 89–97.

Currarini, Sergio, Matthew O. Jackson, and Paolo Pin. 2009. "An Economic Model of Friendship: Homophily, Minorities, and Segregation." *Econometrica* 77(4):1003–45.

Currarini, Sergio, Jesse Matheson, and Fernando Vega-Redondo. 2016. "A Simple Model of Homophily in Social Networks." *European Economic Review* 90:18–39.

DeGroot, Morris H. 1974. "Reaching a Consensus." *Journal of the American Statistical Association* 69(345):118–21.

Dharmalingam, Arunachalam and S. Philip Morgan. 2004. "Pervasive Muslim-Hindu Fertility Differences in India." *Demography* 41(3):529–45.

Dubey, Amaresh, Richard Palmer-Jones, and Kunal Sen. 2006. "Surplus Labour, Social Structure and Rural to Urban Migration: Evidence from Indian Data." *The European Journal of Development Research* 18(1):86–104.

Farooq, Gowhar. 2017. "Politics of Fake News: How Whatsapp Became a Potent Propaganda Tool in India." *Media Watch* 9(1):106–17.

French Jr, John RP. 1956. "A Formal Theory of Social Power." *Psychological Review* 63(3):181.

Golub, Benjamin and Matthew O. Jackson. 2012. "How Homophily Affects the Speed of Learning and Best-Response Dynamics." *The Quarterly Journal of Economics* 127(3):1287–1338.

Granovetter, Mark. 1983. "The Strength of Weak Ties: A Network Theory Revisited." *Sociological Theory* 201–33.

Hegselmann, Rainer, Ulrich Krause, and others. 2002. "Opinion Dynamics and Bounded Confidence Models, Analysis, and Simulation." *Journal of Artificial Societies and Social Simulation* 5(3).

Hoff, Karla and Priyanka Pandey. 2004. *Belief Systems and Durable Inequalities: An Experimental Investigation of Indian Caste*. The World Bank.

Krause, Ulrich. 2000. "A Discrete Nonlinear and Non-Autonomous Model of Consensus Formation." *Communications in Difference Equations* 2000:227–36.

Kumar, Neha. 2014. "Facebook for Self-Empowerment? A Study of Facebook Adoption in Urban India." *New Media & Society* 16(7):1122–37.

Lorenz, Jan. 2007. "Continuous Opinion Dynamics Under Bounded Confidence: A Survey." *International Journal of Modern Physics C* 18(12):1819–38.

McPherson, Miller, Lynn Smith-Lovin, and James M. Cook. 2001. "Birds of a Feather: Homophily in Social Networks." *Annual Review of Sociology* 27(1):415–44.

Mitra, Arup and Mayumi Murayama. 2009. "Rural to Urban Migration: A District-Level Analysis for India." *International Journal of Migration, Health and Social Care*.

Phartiyal S, Ingram D., Patnaik S. 2018. "When a Text Can Trigger a Lynching: WhatsApp Struggles with Incendiary Messages in India." *Reuters UK*.

Shettar, Manoj, Ravichandra Karkal, Anil Kakunje, Rohan Dilip Mendonsa, and VV Mohan Chandran. 2017. "Facebook Addiction and Loneliness in the Post-Graduate Students of a University in Southern India." *International Journal of Social Psychiatry* 63(4):325–29.

Sidhwani, Pranav. 2015. "Spatial Inequalities in Big Indian Cities." *Economic & Political Weekly* 50(22):55–62.

7 Appendix

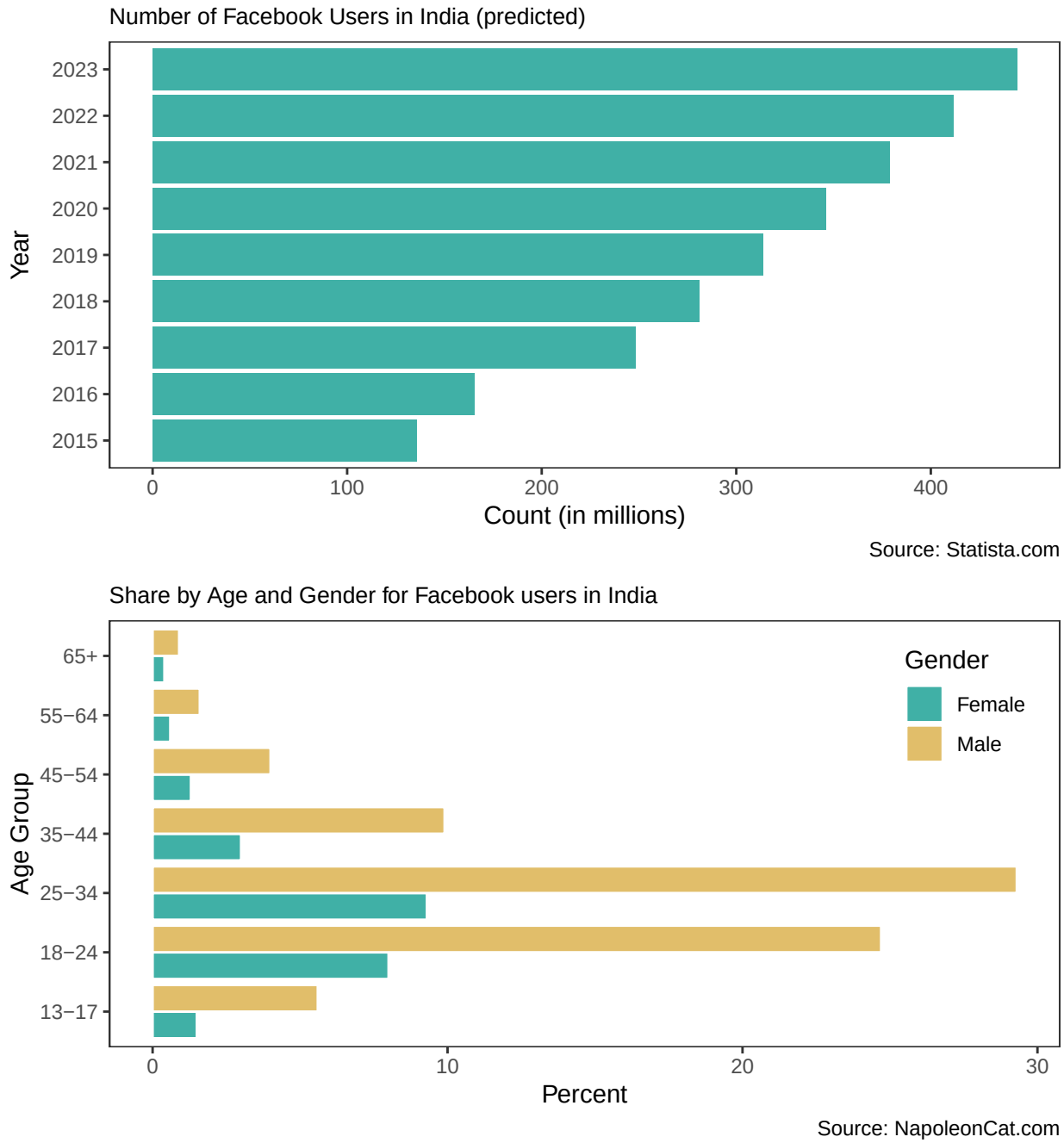

Figure 5: Facebook India User Demographics