Measuring caste homophily in a professional social network

Dharna Prasad[1][0009-0007-1307-8412] and Nisheeth Srivastava[1][0000-0001-9272-8418]

Department of Cognitive Science,
Indian Institute of Technology, Kanpur, India

Abstract. Homophily has been a pervasive phenomenon that has been studied for various characteristics like race and gender in social networks. However, sentiments of selective social association and exclusion have historically also used other demographic vehicles, such as caste. In this paper, we study homophily dynamics of a social network dataset created using LinkedIn connections for Indian university students. We examine these dynamics specifically through the lens of caste, a dominant organizing principle for Hindu society for hundreds, if not thousands of years. Specifically, we estimated the probability of people belonging to specific castes managing to integrate into existing cliques conditional on the caste membership of the cliques, and how these cliques evolve over time. Overall, we find that as the proportion of historically dominant castes dominates a clique, it is less likely for people belonging to historically suppressed castes to join it. A control analysis replacing caste with gender demonstrated null effects, demonstrating the specificity of our observations.

Keywords: social network analysis, caste homophily, time-analysis

1 Introduction

Judgments of social similarity inform all our connections. Homophily refers to the phenomenon where people are more likely to connect with people who are similar to themselves. Studies have measured homophily in different characteristics like race, ethnicity, religion, and occupation (McPherson, Smith-Lovin & Cook, 2001). Homophily in social media connections limits information variability in social networks, which reinforce similar social divides. Users with common attributes are more likely to be friends with each other and form close-knit communities (Mislove and Vishwanath, 2010). Information spreads in the network within these homogenous relationships, thereby producing asymmetries of information that reinforce existing patterns of social dominance (Ertug et al, 2022).

In the US, race is a major factor in social segregation and therefore, race and ethnicity are the primary variables used in the empirical analysis of social networks (Williams et al., 2001). It has been observed that race homophily in networks is different in comparision to gender, often created by preferences rather than base rates. Race homophily is shaped by structural divide and group-size differences, whereas for gender, both are present in similar numbers (McPherson et al. 2001).

In contrast, in India, societal structure and segregation has been dominated by caste divisions extending far into antiquity (Ambedkar, 1987). “Jati” is the vernacular word for caste in India, used to define the cohesive group one belongs to. The word 'jati' etymologically means 'something into which one is born’. Caste is the English translation of ‘Jati’ and refers to a fixed social group one belongs to in the caste structure. Historically, members of suppressed castes were excluded from resources like education, health facilities, clothing by dominant ‘upper’ castes based on claims of retaining 'purity'. Strikingly, extremely strict caste-based endogamy has been a feature of Indian society, attested clearly by recent archaeogenetic evidence, ranging from between 30 to 100 generations into the past (Reich et al., 2009).

Even in modern India, a person born in a particular caste takes on a caste-associated surname. The caste identity of people takes precedence over their individuality and becomes their source of their judgement (Guru, 2014). An essential characteristic of caste lies in persistent stigmatization of the lower castes by higher caste groups. High castes tend to show more extreme tendencies towards homophily than lower caste groups (T. Davidson, 2018). Since India's independence, laws have been made to strengthen the dignity of people irrespective of their social background. However, India still exhibits prominent caste structures, since it is difficult to break through rigid hierarchies.

A study of caste dynamics among Indian Politicians on Twitter, showed that higher caste politicians are likely to express stronger influence and are structurally advantageous positions on social media compared to lower caste Politicians (Vaghela et al., 2022). Societies that build hierarchical structures in the real world are likely to follow similar structures in online interactions (Smith et al. 2001), an insight empirically substantiated by multiple empirical studies of online social networks (Bisgin, Agarwal & Xu, 2010).

In this paper, we empirically examine the dynamics of homophily on the LinkedIn professional social network from the lens of caste, the distinctive sociological determinant of social association in India.

2 Methods

We create our data of users on LinkedIn that are connected to each other by commenting on each other's posts. Our dataset is limited to an Indian university students, since information extraction from LinkedIn is only possible realistically for common members of specific LinkedIn communities without bespoke APIs. We searched for students with this university as their alma mater for the years 2022-24, to curate a list of students, for which we extracted their comment interaction on posts made by them. We observed a subset of users engage in passive social media behaviors like scrolling through and viewing others' profiles without initiating active interactions. Active interaction is demonstrated through comments and posts. We retain only those users who actively interact with others. Our analysis dataset consists of users, and all the people who commented on user’s posts, till the time of our data collection (January 2025).

We structured our data collection in a way that targets a specific user’s (A’s) post and collects all the users who have commented on that post. In this way, for each post authored by a user, we extract names of all the commenters who have interacted with it. It forms a dataset (the Table uses fictional names) that looks like:

poster_name	commenter_name
Apple Sharma	J. Mathur
Apple Sharma	Apoorv Kumar
Ram G.	Shiv Kapur

We assume that comment relationship reflects a bidirectional interaction, meaning both users are socially connected. Using this, we create an undirected edge A-B. Any self-loops of people who have replied to themselves are removed, as it does not contribute to the network. Using the python networkx library, we create a network of users on LinkedIn that are connected to each other by commenting on each other's posts. People are represented as nodes, and edges between them are inferred from their comment relationships. Descriptive statistics for the dataset are given in Table 1.

Network Descriptive Statistics				
	Nodes	Edges	Nodes in LCC*	Edges in LCC
Dataset1	55768	66868	49880	61596
Dataset2	27742	31709	23639	28482
LCC = Largest Connected Component*				
Dataset 2 is separately collected with timestamps				

Table 1. Dataset Information

Time Analysis Data Collection

For time analysis, we collected data again to collect timestamps for every comment interaction. In our network data for time analysis, we have interaction of Indian university students over LinkedIn, refer to Table 1 for descriptive statistics for the same. LinkedIn has been around for years now, and people have varied usage of LinkedIn. One may stay on LinkedIn while looking for jobs, and not otherwise. This stands as a confound for our time analysis, as some people may use it continuously overtime, and some may not. For the analysis, we assume that cliques-built overtime persists throughout. We organize our data on time points scaling in months [3 months, to 3 years]. In these timestamps, we look for subgraphs, as in, cliques formed and their category composition. We used a survival analysis framework, to understand which type of people persists through groups. We use mixed effects Cox Regression model to measure the probability of an SC joining a clique at the next timestamp, given that the subject has survived with certain UC proportion till that time point.

Dictionaries: Caste and gender

The next step in our analysis involves categorizing nodes of our network based on characteristics like caste identity or gender. We attempted to categorize our data for caste identities using the naming conventions. To identify caste identities from names is a difficult task, but generalization is present in everyday usage of names and their caste identities. As noted by Shrestha (2000) - "'Each caste has characteristic surnames by

which members can be identified and placed in the hierarchy; thus, the surnames indicate to others within the overall caste system the degree of deference or authority a Newar should have"

Names not only act as a point of identification but also signal for geographical and social identity. Social characteristics are often indicated via names. Building on the premise that names encode social attributes beyond simple identification (Brennen, 2000), we use them to identify homophily patterns within a social media network. In India, names are a significant indicator of caste. Personal and family names code the caste (Jati) identities. Post independence, thousands of caste groups were administered into the designation of Scheduled Castes, Scheduled Tribes, and Other Backward Castes. In daily life, one can predict ethnicity, caste, or economic wealth from the title of a person.

A similar case of race identification through names is studied in the US. Bertrand and Mullainathan (2004) found that White names were more likely to get a call for an interview than Black names. Sandeep Bhupatiraju (2024) studied social identity impact on judicial processes, inferring caste status from Surnames using machine learning. They found that caste neutral litigants are likely to choose lawyers from similar communities, and often disadvantaged in judiciary procedures like case dismissals, etc. Another study, measuring caste homophily in management students, found that elite groups showed stronger homophily to alike ties, whereas Lower status groups sought more ties, and resulted with homophilous withdrawal [1].

Building on such earlier efforts, we use UPSC ranks lists to extrapolate caste identities using names. Since these lists explicitly mention categories applied under, we use them to categorize our dataset. Using this, we create dictionaries of people's surnames that belong to each Category – Upper Castes [UC], Suppressed Castes [SC], Gray. Since some names may be identified in both categories as they could caste neutral names, these are categorized in gray area (Jayaraman, 2005). We use these dictionaries to assign categories to people in our network, based on similarity and identifying surnames from our compiled dictionaries. From our dictionary categorization, we found following categories:

UC – 16372; SC - 10329; Gray - 5598; Unclassified- 23478.

We indicate surnames that we could not categorize as Unclassified. Such surnames either could be shortform, unknown or belonging to other religions, which may or may not directly follow caste system.

3 Results

Clique Analysis through a Caste Lens

We have built an undirected network of user interactions on LinkedIn, now we move on to the study of sub-graphs formed in our network. Using networkx, we identify cliques in the network. A clique is defined as a subgraph of nodes that are completely connected to each other.

[1] Homophilous withdrawal refers to less stable ties with outside group, leading one to go back to their own social circle.

We want to see the dissimilar nodes in a clique, and how they are integrated as the size of the clique grows. We have three categories in the data - UC, SC, and Gray. The least found category in a group is referred to as the Minority. There are multiple cliques formed in different sizes. To locate the Minority fraction at a specific clique size, we take the average minority fraction found in all cliques of a particular size. These averaged minority fractions are plotted against their size. In fig 1, we observe the trend that as clique size is increasing, the minority fraction is decreasing. This means that as the size of a clique grows, the diverse category of people is diminishing, and the clique is being populated by a similar category of nodes, thus demonstrating evidence of generic caste homophily in the network.

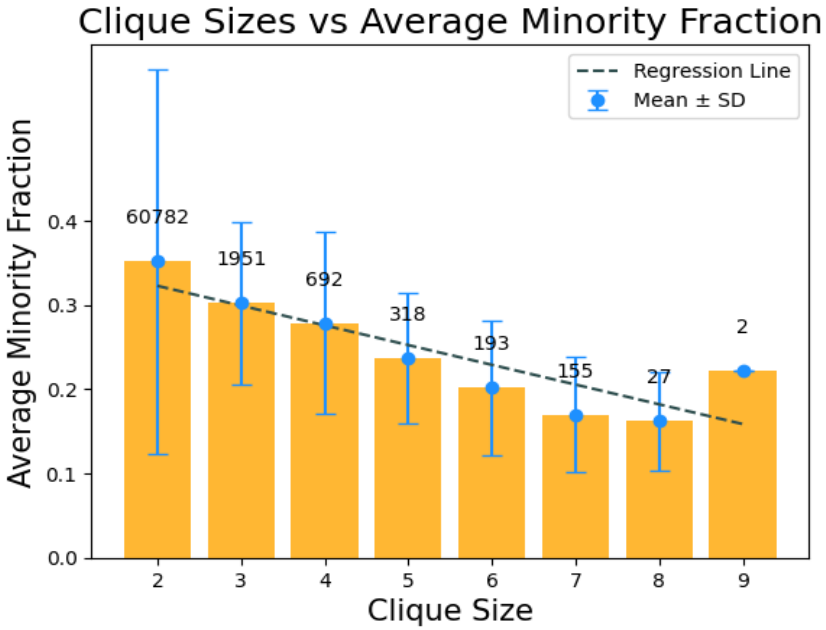

Fig. 1. Average minority fraction against Clique Size, with regression line indicating their relationship. On top, are the number of cliques found at each size, with mean and SD for minority fractions at each size.

We conducted a simple regression analysis using OLS to test the relationship between clique sizes and minority fractions. The model was significant, $F = 18.30$, $p = 0.005$, and explained approximately 75.3% of the variance in y ($R^2 = 0.753$). This indicates a strong, statistically significant negative association between clique sizes and minority fractions.

The next question we asked was how the categories are affected by growing clique sizes. For this, we look at the average SC fraction and average UC fraction at each clique size. Figure 2 shows the average SC fraction plotted against growing clique sizes. Interestingly, we observe that as clique sizes increase, the average fraction of SC is

decreasing, whereas it is increasing for the average UC fraction. Thus, homophily of UC dominant nodes increases as the size of the clique grows, but not for SC dominant nodes.

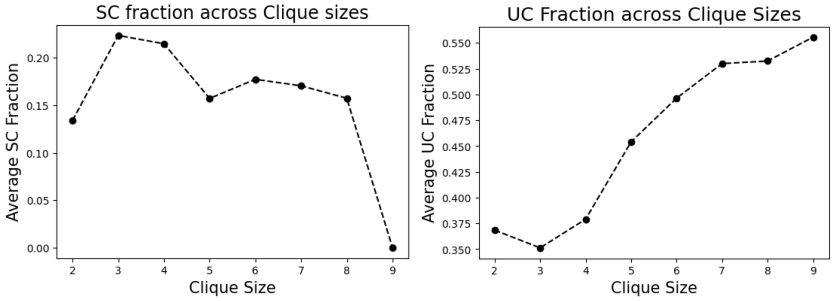

Fig. 2. (A) Average SC fraction vs clique size. (B)Average UC fraction vs clique size.

These distinctive stratified results suggested that the overall caste homophily seen in the network may be significantly driven by social preferences for homophilic association in UC cliques. This mechanism would offer one possible explanation for the greater UC fraction in larger cliques compared to smaller ones.

Temporal Analysis of Homophily Dynamics

Our primary network analysis revealed greater UC presence in larger cliques than smaller ones, but not for SC. While UC caste homophily could offer one possible explanation for this result, greater discoverability of the majority class (UC) in recommendation feeds could offer an alternative, and more innocuous explanation. To discriminate between these two possibilities, we conducted a temporal analysis in which we estimated the probability of a SC joining a clique is based on the proportion of UC in the clique prior to their joining based on timestamps of interactions in our second dataset.

This second dataset consisted of repeated measurements of clique memberships over time. Given this longitudinal structure and time-to-event outcomes, we applied mixed-effects Cox regression used for survival analysis, to account for within-clique correlations and time-varying risk. In survival analysis terminology, we study the event of SC/UC joining a clique, considering the proportion of clique prior to the event. We formalized the model_1 as: Surv(time, status_SC) ~ UC_frac + (1 |clique_members) and observed 210 events, in 7797 observations. The second Model examines the converse relationship, the rate of UC joining a clique, considering the proportion of SC population prior to joining, Table 2 presents the results from our mixed effects cox regression models.

Mixed Effects Cox Regression Models

Characteristic	exp(Beta)	95% CI	p-value
Model 1: UC_frac (SC join)			
UC_frac	0.17	0.10, 0.27	<0.001
Model 2: SC_frac (UC join)			
SC_frac	0.34	0.22, 0.54	<0.001

Abbreviation: CI = Confidence Interval

Table 2. Model 1 examines the rate of SC joining a clique with prior population of UC (83% decrease), and Model 2 examines the rate of UC joining a clique with prior clique proportion of SC (66% decrease).

The results indicate that the proportion of UC in a clique decreases the probability of SC joining the clique by 83% ($\exp(\beta) = 0.17$, 95%CI: 0.10-0.27, $p < 0.001$). Standard deviation (0.0090) and variance (8.17e-05) of the random effects come out to be very small, indicating that most of the variation in joining rate is explained by proportion of UC rather than differences between cliques. Whereas prior population of an SC person, decreases the likelihood of a UC joining by 66% ($\exp(\beta) = 0.34$, 95%CI: 0.22-0.54, $p < 0.001$). These results produce stronger support for the caste homophily explanation than base rate effects as an explanation for the original result. Simply, we find that a Suppressed caste is less likely to join a UC dominated clique by 83%, whereas probability of a UC joining a SC dominated cliques stands by 66%.

Clique and Temporal analysis through Gender Lens

To test the specificity of our analysis, we performed a similar sub analysis on the network [Dataset2] from the lens of gender instead of caste. We categorized all nodes in the network by gender, using dictionaries for common male and female suffixes to assign gender. We use the second dataset for gender analysis. Of the total nodes, 15,411 were classified as male, 6845 as female, and 4355 as unknown. This network is used to study the distribution of gender in its sub-graphs. We calculate average male and female fractions in each clique and plot them against clique sizes. In Figure 3, we observe the trends that for gender the fraction remains similar for male or female. These trends are different from the trends observed in caste categorization and reflect differences in homophily between the two characteristics. We fit simple regression to study the effect of each gender on clique size. We observe that the male fraction shows a strong, statistically significant relationship with clique size ($R^2 = 0.781$, $p = 0.019$), whereas female fraction shows no meaningful relationship ($R^2 = 0.028$, $p = 0.751$). Male fraction of cliques can explain portion of clique sizes, whereas female fraction seems to not have an effect.

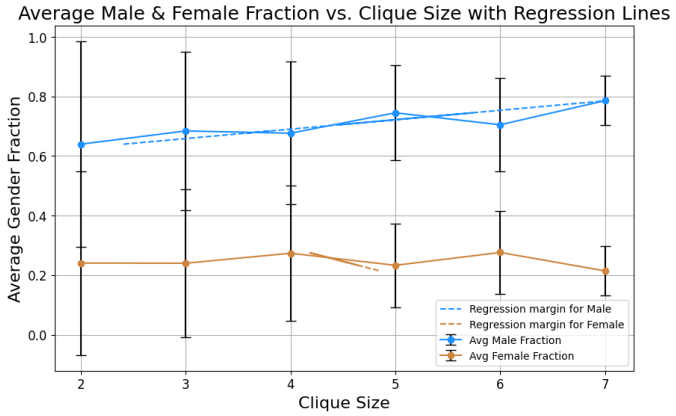

Fig. 3. Average male and female fraction in Cliques by size, with error bars indicating SD.

We conduct similar survival analysis on overtime data to examine how the gender proportion of a clique affects the likelihood of a particular gender joining the clique. We formalized the model_1: Surv(time, status_female_join) ~ male_frac + (1 |clique_members and model_2 as: (Surv(time, status_male) ~ female_frac + (1 | clique_members). The results are presented in Table 3. We observe 90% decrease in the likelihood of female joining cliques with higher male proportion, whereas there is 55% decrease in male joining a clique that has majority female proportion. Considering previous clique analysis, it seems interesting that male proportion does not structurally affect clique sizes but does affect the interaction of females in entering cliques.

Mixed Effects Cox Regression Models

Characteristic	exp(Beta)	95% CI	p-value
Model 1: male_frac (female join)			
male_frac	0.10	0.08, 0.13	<0.001
Model 2: Female_frac (male join)			
female_frac	0.45	0.36, 0.55	<0.001

Abbreviation: CI = Confidence Interval

Table 3. Model 1 assesses the rate of Female joining the group with prior proportion of Males in the cliques, and Model 2 looks at the rate of Male joining a group with prior female proportion in cliques

4 Discussion

Our study of LinkedIn of Indian university students revealed that online platforms replicate homophily of caste identities for UC, but not for SC. These results are consistent with studies that have indicated strong caste homophily in organizations and rural India, and personal experiences of people. (T, Davidson (2018); Bhardwaj et al. (2021). Our results observed that the present of any minority in a clique affects the size of the clique. This is consistent with Blau's theory of social structure that suggests that the relative size of groups (i.e., minority/majority status) affects interaction patterns and group boundaries (Blau, 1977).

Our time analysis shows that over time, it becomes difficult for a person from a suppressed caste to enter an already cohesive group of upper caste members, consistent with exercise of deliberate social preference rather than base rate or interface effects on clique membership. These results support upper caste homophily as a powerful explanation for the organization of social structures in Indian society, consistent with existing sociological and biological evidence pointing in the same direction (Reich et al., 2009). Paralleling our caste analysis, we also found that females, but not males, find it harder to assimilate into cliques dominated by the other gender.

Our results are consistent with evidence for demographic homophily in other elite online contexts. For example, on Reddit's r/news forum, Monti et al. (2023) find that demographic similarity (e.g., age, income) predicts interaction more strongly than ideological agreement, even within high-engagement threads. Similar structures have been found in enterprise communication networks, where high-status individuals form dense "rich club" subnetworks based on role and demographic homogeneity (Dong et al., 2014).

The medium of our investigation, however, has important restrictions that are necessary to acknowledge. It was a difficult task to collect data from LinkedIn, since restrictions have been put on public data access on social media. Our categorization of caste identity may be severely restricted in its full representation. Although we used simple categorization of Reserved and Unreserved categories from UPSC data, such categorization may not capture the complexities of caste in India. We only analyze the influence of SC and UC in our data, ignoring other ethnicities which may be constraining our results. We did not focus on the gray category, which is meaningful in depicting the grayness in the caste of people with ambiguous surnames. In future analysis, the gray category could be useful for looking at middle-sized groups' influence on networks. Finally, our definition of interactions as being bijective with relationships is fairly arbitrary, and alternative specifications may well have more ecological validity.

References

1. Ambedkar, B. R. (1987). *Dr. Babasaheb Ambedkar Writings and Speeches: Vol. 4/Vol. 9*. Government of Maharashtra.
2. Banerjee, A., Chandrasekhar, A. G., Duflo, E., & Jackson, M. O. (2013). The diffusion of microfinance. *American Economic Review, 103*(6), 1892–1926.
3. Bertrand, M., & Mullainathan, S. (2004). Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review, 94*(4), 991–1013.

4. Bhardwaj, A., Mishra, S. K., Qureshi, I., Kumar, K. K., Konrad, A. M., Seidel, M.-D. L., & Bhatt, B. (2021). Bridging caste divides: Middle-status ambivalence, elite closure, and lower-status social withdrawal. *Journal of Management Studies, 58*, 2111–2136.
5. Bisgin, H., Agarwal, N., & Xu, X. (2010, August). Investigating homophily in online social networks. In *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* (Vol. 1, pp. 533–536). IEEE.
6. Blau, P. M. (1977). *Inequality and heterogeneity: A primitive theory of social structure*. Free Press.
7. Brennen, T. (2000). On the meaning of personal names: A view from cognitive psychology. *Names, 48*(2), 139–146.
8. Centola, D. (2011). An experimental study of homophily in the adoption of health behavior. *Science, 334*(6060), 1269–1272. https://doi.org/10.1126/science.1207055
9. Currarini, S., Jackson, M. O., & Pin, P. (2009). An economic model of friendship: Homophily, minorities, and segregation. *Econometrica, 77*(4), 1003–1045.
10. Davidson, T. (2018). Variation in caste homophily across villages and contexts in rural India. *Unpublished manuscript*.
11. Deshpande, A., & Sharma, S. (2013). Entrepreneurship or survival? Caste and gender of small business in India. *Economic and Political Weekly, 48*(28), 38–49.
12. Dong, Y., Yang, Y., Tang, J., & Chawla, N. V. (2014). Inferring user demographics and social strategies in enterprise social networks. *Proceedings of the 22nd International Conference on World Wide Web*, 739–750.
13. Ertug, G., Brennecke, J., Kovács, B., & Zou, T. (2022). What does homophily do? A review of the consequences of homophily. *Academy of Management Annals, 16*(1), 38–69.
14. Jayaraman, R. (2005). Personal identity in a globalized world: Cultural roots of Hindu personal names and surnames. *The Journal of Popular Culture, 38*(3), 476–490.
15. Kossinets, G., & Watts, D. J. (2006). Empirical analysis of an evolving social network. *Science, 311*(5757), 88–90.
16. Kossinets, G., & Watts, D. J. (2009). Origins of homophily in an evolving social network. *American Journal of Sociology, 115*(2), 405–450.
17. McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology, 27*, 415–444.
18. Monti, C., D'Ignazi, J., Starnini, M., & De Francisci Morales, G. (2023, April). Evidence of demographic rather than ideological segregation in news discussion on Reddit. In *Proceedings of the ACM Web Conference 2023* (pp. 2777–2786).
19. Patel, K. (2017). What is in a name? How caste names affect the production of situated knowledge. *Gender, Place & Culture, 24*(7), 1011–1030.
20. Reich, D., Thangaraj, K., Patterson, N., Price, A. L., & Singh, L. (2009). Reconstructing Indian population history. *Nature, 461*(7263), 489–494.
21. Shrestha, U. (2000). Changing patterns of personal names among the Maharjans of Katmandu. *Names, 48*(1), 27–48.
22. Vaghela, P., Kommiya Mothilal, R., Romero, D., & Pal, J. (2022). Caste capital on Twitter: A formal network analysis of caste relations among Indian politicians. Proceedings of the ACM on Human-Computer Interaction, 6(CSCW1), Article 80.
23. Williams, D. R., & Collins, C. (2001). Racial residential segregation: A fundamental cause of racial disparities in health. *Public Health Reports, 116*(5), 404–416.
24. Wimmer, A., & Lewis, K. (2010). Beyond and below racial homophily: ERG models of a friendship network documented on Facebook. *American Journal of Sociology, 116*(2), 583–642.