https://doi.org/10.1038/s44260-024-00009-9

Networks and identity drive the spatial diffusion of linguistic innovation in urban and rural areas

Check for updates

Aparna Ananthasubramaniam[1], David Jurgens[1,2] & Daniel M. Romero[1,2,3]

Cultural innovation (e.g., music, beliefs, language) tends to be adopted regionally. The geographic area where innovation is adopted is often attributed to one of two factors: (i) speakers adopting new behaviors that signal their demographic identities (i.e., an *identity* effect), or (ii) these behaviors spreading through homophilous networks (i.e., a *network* effect). In this study, we show that network and identity play complementary roles in determining where new language is adopted; thus, modeling the diffusion of lexical innovation requires incorporating both network and identity. We develop an agent-based model of cultural adoption, and validate geographic properties in our simulations against a dataset of innovative words that we identify from a 10% sample of Twitter (e.g., fleeky, birbs, ubering). Using our model, we are able to directly test the roles of network and identity by comparing a model that combines network and identity against simulated network-only and identity-only counterfactuals. We show that both effects influence different mechanisms of diffusion. Specifically, network principally drives spread among urban counties via weak-tie diffusion, while identity plays a disproportionate role in transmission among rural counties via strong-tie diffusion. Diffusion between urban and rural areas, a key component in innovation spreading nationally, requires both network and identity. Our work suggests that models must integrate both factors in order to understand and reproduce the adoption of innovation.

From new technologies[1,2], to religious beliefs[3,4] to popular music[5,6] and memes on social media[7,8], innovation is often adopted regionally within the USA (e.g., in the Deep South or the Mid-Atlantic)[9,10]. For instance, new words are often used in geographic areas that reflect their social, cultural, and historical significance[11,12]. In fact, many social science disciplines (e.g., sociology, anthropology, linguistics, cultural, and social geography) use linguistic variables as a proxy for culture change[13–16], because shifts in culture often result in language change, and conversely, using new language sometimes signals adoption of new worldviews[17–19]. Specifically, researchers often use the geographic regions where new language is adopted to test putative mechanisms of diffusion[20–23]: To falsify a hypothesized mechanism, one could show that it does not predict where speakers would adopt a new word.

Existing mechanisms often fail to explain why cultural innovation is adopted differently in urban and rural areas[24–26]. Urban centers are larger, more diverse, and therefore often first to use new cultural artifacts[27–29]. Innovation subsequently diffuses to more homogenous rural areas, where it starts to signal a local identity[30]. Urban/rural dynamics in general, and diffusion from urban-to-rural areas in particular, are an important part of why innovation diffuses in a particular region[24–27,29–31], including on social media[32–34]. However, these dynamics have proven challenging to model, as mechanisms that explain diffusion in urban areas often fail to generalize to rural areas or to urban-rural spread, and vice versa[30,31,35].

Spatial properties of diffusion are often hypothesized to be the result of one of two mechanisms: the performance of demographic *identity* (henceforth referred to simply as identity) or the diffusion of innovation through a homophilous *network* (henceforth, network)[10,30,31]. On one hand, speakers may adopt language that allows them to perform their demographic identity—using certain words to signal what identities they hold (e.g., saying “pop” instead of “soda” to sounds Midwestern)[13,36,37]. For instance, mechanisms like strong-tie diffusion suggest that demographically similar speakers (often connected by strong, or close, ties) influence each others’ adoption[38–40], explaining geographic variation as the byproduct of spatial assortativity in personal characteristics[11,35,41]. On the other hand, language

[1]School of Information, University of Michigan, 105 S State St, Ann Arbor, MI, 48109, USA. [2]Computer Science and Engineering Division, Electrical Engineering and Computer Science Department, University of Michigan, 2260 Hayward St, Ann Arbor, MI, 48109, USA. [3]Center for the Study of Complex Systems, University of Michigan, 500 Church St, Ann Arbor, MI, 48109, USA. ✉e-mail: akananth@umich.edu

regions may also be the result of network homophily—or the tendency for similar individuals to be connected in the social network (e.g., Michiganders tend to have ties to other Michiganders, Democrats to other Democrats)[28,40,42,43]. The amount of homophily in a network has been shown to determine both the extent of diffusion[44,45], as well as specific geographic properties of cascades[46]. For instance, mechanisms like weak-tie diffusion suggest that new words tend to diffuse via the network, where weak ties, or more distant relationships, increase a word's exposure[43,47,48]; via this mechanism, geographically and demographically homophilous ties allow language regions to emerge[49–52]. As an example, let's assume the phrase "no human is illegal" is more likely to be used in politically left-leaning states. Under the identity effect, this adoption geography is expected because using the phrase makes a speaker sound like a Democrat, and, therefore, it would likely diffuse in areas where many Democrats live and choose to use it[35]. Under the network effect, the phrase is thought to spread in left-leaning states because, once some Democrats start using it, their (largely Democratic) friends and neighbors start repeating it.

Existing theory tends to focus on either network or identity as the primary mechanism of diffusion. For instance, cultural geographers rarely explore the role of networks in mediating the spread of cultural artifacts[53], and network simulations of diffusion often do not explicitly incorporate demographics[54]. Even within fields that acknowledge both network and identity as drivers of diffusion (e.g., sociology theories of diffusion or variationist sociolinguistics), any given model of adoption is often either identity-centered or network-centered, rather than offering an explanation of diffusion that connects the two[35,55–58]. Urban/rural dynamics are not well-explained using these network- or identity-only theories; in particular, in some cases, identity-only frameworks designed to model rural adoption do not explain urban diffusion[30], while some network-only models capture urban but not rural dynamics[31]. However, a framework combining both of these effects may better explain how words spread across different types of communities[59].

In this study, we test whether network and identity play complementary roles in creating key spatial properties of lexical diffusion. Specifically, we hypothesize that network tends to drive weak-tie diffusion between urban counties, while identity promotes strong-tie diffusion between rural counties. Testing our hypothesis requires comparing a combined network + identity model of diffusion to network-only and identity-only counterfactuals—and since network and identity are often correlated[50], we cannot empirically observe these baselines. Instead, we develop an agent-based model, inspired by cognitive and social theory, to model the spread of new words through a network of speakers. Using agent-based models allows us to simulate the required counterfactuals and, therefore, directly test how network and identity interact[60]. Our simulations are validated using large-scale empirical data we curate, including a registry of new words on the microblog site Twitter (now known as 𝕏) and the network and demographic identities of users on the site.

We find evidence supporting our hypothesis and, therefore, that key properties of linguistic diffusion—both the geographic regions that new words spread to and the spatiotemporal pathways through which they diffuse—are better approximated by network and identity together than by either one individually. Furthermore, urban/rural heterogeneity is an emergent property of our model: differences between urban and rural counties are present when taking network and identity into account, even though we do not explicitly model them. We conclude that models omitting either network or identity are missing a crucial dynamic in the adoption of innovation and drawing incomplete conclusions about the underlying diffusion process.

Methods

We develop an agent-based model to evaluate the roles of network and identity in the spatial patterns of cultural diffusion. To realistically model the adoption of innovation, our formulation draws heavily from social and cognitive theory, and underlying assumptions are empirically derived[61–64]. Our model simulates the diffusion of a new word w. The model begins with a set of initial adopters introducing the word to the lexicon (section "New words and initial adopters"), and spreads across a directed network of n agents $\{j\}_{j=1}^{n}$ (section "Network" and section "Agent identity"). The new word connotes a particular identity $\boldsymbol{\Upsilon}_w$ that is assigned based on the identities of its early users (section "Word identity"). In our simulations, the word continues to spread through the network over several subsequent timesteps (section "Diffusion"). Agents are exposed to the word when a network neighbor uses it. Agents are more likely to use the word if it signals an identity congruent with their own and if they were recently exposed by network neighbors with similar identities. We fit the model's free parameters to empirical data about each word's diffusion (section "Parameters and trials"), and compare how well this full model reproduces properties of empirical trials (section "Model evaluation" and section "Testing the hypotheses") relative to network- and identity-only counterfactuals (section "Simulated counterfactuals"). See Supplementary Methods 1.2 for the full set of model equations and Supplementary Methods 1.3 for information about parameters and how they are inferred. Our model's limitations, along with our attempts to address them, are listed in the Supplementary Discussion. Although we test our model against the diffusion of linguistic innovation (section "Hypotheses"), its formulation is sufficiently general to describe the adoption of other cultural innovations.

New words and initial adopters

We simulate the diffusion of widely used new words originating on Twitter between 2013 and 2020. Starting from all 1.2 million non-standard slang entries in the crowdsourced catalog UrbanDictionary.com, we systematically select 76 new words that were tweeted rarely before 2013 and frequently after (see Supplementary Methods 1.41 for details of the filtration process). Consistent with prior studies of online innovation[65–69], the 76 new words in our study include terms describing popular culture phenomena (e.g., fanmix, sweaties), phonologically-motivated orthographical shifts (e.g., bawmb, whatchoo), part-of-speech changes (e.g., ubering, lebroning), abbreviations (e.g., ihml, profesh), concatenations (e.g., amaxing, sadboi), and even new coinages (e.g., gwuap, fleeky) (Supplementary Table 3 has more examples). These words often diffuse in well-defined geographic areas that mostly match prior studies of online and offline innovation[23,69] (see Supplementary Fig. 7 and Supplementary Methods 1.4.4 for a detailed comparison).

Each run of our model simulates the diffusion of one of these 76 words. The set of final adopters is often highly dependent on which users first adopted a practice (i.e., innovators and early adopters)[70], including the level of homophily in their ties and the identities they hold[71,72]. Therefore, we seed the model with a set of empirical early adopters. Each simulation's initial adopters are the corresponding word's first ten users in our tweet sample (see Supplementary Methods 1.4.2). Model results are not sensitive to small changes in the selection of initial adopters (Supplementary Methods 1.7.4).

Network

Patterns in the diffusion of innovation are often well-explained by the topology of speakers' social networks[42,43,73–75]. Therefore, the word in our model diffuses through a network of agents. Nodes (agents) and edges (ties) in this network come from the Twitter Decahose, which includes a 10% random sample of tweets between 2012 and 2020. Agents in our model correspond to Twitter users in this sample who are located in USA. We draw an edge between two agents i and j if they mention each other at least once (i.e., directly communicated with each other by adding "@username" to the tweet), and the strength of the tie from i to j, w_{ij} is proportional to the number of times j mentioned i from 2012 to 2019[76,77]. The edge drawn from agent i to agent j parametrizes i's influence over j's language style (e.g., if w_{ij} is small, j weakly weighs input from i; since the network is directed, w_{ij} may be small while w_{ji} is large to allow for asymmetric influence). Although Twitter users are exposed to content from more users than they reciprocally mention (e.g., unreciprocated ties, users they follow, public tweets), this network is particularly relevant to our study; prior research has shown that the mention network captures edges likely influential in information diffusion[78], and reciprocal ties are often responsible for the diffusion of lexical items[79] and better predict properties of cascades[80]. Moreover, reciprocal ties are more likely to be structurally balanced and have stronger triadic closure[81], both of which facilitate information diffusion[82].

This directed network has nearly 4 million nodes and 30 million edges; the network evidences homophily (higher than expected levels of assortativity along all modeled aspects of identity) and exhibits some clustering within geographically localized regions as well as some clustering across regions (Supplementary Figs. 2–4). The network also exhibits expected patterns in urban and rural tie strength. Consistent with prior studies of urban and rural areas[30,83], ties between two urban counties tend to be weak ties (less demographic similarity and lower edge weight), while ties between two rural counties tend to be strong ties (more demographic similarity and higher edge weight) (Supplementary Figs. 18, 19). As expected, demographic similarity and edge weight are correlated: ties with lower edge-weight w_{ij} tend to share fewer demographic similarities than edges with higher weight (Supplementary Table 6).

Model results are robust to modest changes in network topology, including the Facebook Social Connectedness Index network (Supplementary Methods 1.7.1)[84] and the full Twitter mention network that includes non-reciprocal ties (Supplementary Methods 1.7.2).

Agent identity

An individual often adopts innovation that signals their affiliation with some identity[37,85–87]. In our model, area demographics are proxies for each agent's probable identity. Note that, although the term "identity" typically refers to how someone identifies along a range of markers[88], our paper models solely demographic aspects. Agents are characterized by $D = 5$ categories shown to be important to language style: (i) location within USA[21,89,90], (ii) race/ethnicity[91–94], (iii) socioeconomic status measured via income level, educational attainment, and workforce participation[47,95,96], (iv) languages spoken[97–99], and (v) political affiliation[14,100]. Each category is parametrized by several related registers (e.g., for political affiliation, "registers" are Democrat, Republican, and Third Party), for a total of $d = 26$ registers.

We infer each agent's location from their GPS-tagged tweets, using Compton et al. (2014)'s algorithm[101]. To ensure precise estimates, this procedure selects users with five or more GPS-tagged tweets within a 15-km radius, and estimates each user's geolocation to be the geometric median of the disclosed coordinates (see Supplementary Methods 1.1.2 for details). By using conservative thresholds for frequency and dispersion, this algorithm has been shown to produce highly precise estimates of geolocation. Since Twitter does not supply demographic information for each user, agent identities must be inferred from their activity on the site. Automated demographic recognition tools often use network ties (or posts with mentions) as features, which would preclude independent measures of identity and network, and there are some debates around the methodological soundness and ethical acceptability of these methods[102–104]. Instead, we estimate each agent's identity based on the Census tract and Congressional district they reside in refs. 105,106. Similar to prior work studying sociolinguistic variation on Twitter[12,107], each agent's race/ethnicity, SES, and languages spoken correspond to the composition of their Census Tract in the 2018 American Community Survey. We also represent each agent's political affiliation using their Congressional District's results in the 2018 USA House of Representatives election. Since Census tracts are small (population between 1200 and 8000 people) and designed to be fairly homogeneous units of geography, we expect the corresponding demographic estimates to be sufficiently granular and accurate, minimizing the risk of ecological fallacies[108,109]. Due to limited spatial variation (Supplementary Methods 1.1.4), age and gender are not included as identity categories even though they are known to influence adoption. However, adding age and gender (inferred using a machine learning classifier for the purposes of sensitivity analysis) does not significantly affect the performance of the model (Supplementary Methods 1.7.3).

Since an agent may identify with each identity register to a different degree[37,110] and in order to capture spatial variation, each register of an agent's identity Υ_j is represented as a value in the interval [0, 1] (e.g., in a district where 61% voted Republican and 39% Democrat, the Republican identity is represented by 0.61 and Democrat identity as 0.39, instead of the majority identity of 1 and 0, respectively), so $\Upsilon_j \in [0, 1]^d$. Even though this procedure may underestimate some variation in demographics (e.g., in the example above, a Republican and a Democrat in the district are both represented with political identities of (0.61, 0.39)), our estimation strategy captures the spatial variation in identities that are hypothesized to drive geographic patterns in language diffusion. In particular, we did not randomly assign identities within Census tracts in order to avoid obscuring homophily in the network (i.e., because random assignment would not preferentially link similar users).

Word identity

Cultural innovation can be used to signal different aspects of an agent's identity[111–113]. Each word may provide information about one or more of the identity categories like location, race, etc.[88]; for each word, we denote the relative importance of each category with weight vector $\mathbf{v}_w \in [0, 1]^D$. Unlike agent identity, words often connote affiliation with a specific register of identity (e.g., in Eckert 2000, high schoolers may associate with multiple social groups, but each linguistic variable signals membership to a particular group[114]). Therefore, word identities in our model are binary (i.e., a word either signals a given register of identity or it doesn't), and we model word identities distributed in $\Upsilon_w \in \{0, 1\}^d$ unlike agents' identities in $\Upsilon_j \in [0, 1]^d$.

A word's identity is often enregistered based on the demographics of a small number of its early adopters[110], signaling that these speakers identify with certain registers of identity. For instance, if the initial adopters tend to come from disproportionately Republican, African American, French-speaking areas like Louisiana, the word signals this demographic identity: specifically, $v_w = \frac{1}{3}$ for the dimensions corresponding to the political affiliation, race, and language categories; $\Upsilon_w = 1$ for the dimensions corresponding to the Republican political affiliation, African American race, and French language registers; and other entries of both $\mathbf{v}_w$ and Υ_w are 0 (see Supplementary Methods 1.2.2–1.2.3 for a more formal description). Agent identities remain unaltered by a word's enregisterment. During the process of enregisterment, both online and offline, words often quickly develop a "stereotypic indexical value," or universal understanding of the identity signaled by the word shared by all speakers and conveyed through context[71,115,116]. Therefore, a word's identity is assigned based on the word's first ten adopters.

Diffusion

After the initial adopters introduce the innovation and its identity is enregistered, the new word spreads through the network as speakers hear and decide to adopt it over time. In order to appropriately model the diffusion of language[18], adoption is usage-based (i.e., agents can use the word more than once and adoption is influenced by frequency of exposure)[117] and the likelihood of adoption increases when there are *multiple* network neighbors using it[118]. Although we present a model for lexical adoption on Twitter, the cognitive and social processes on which our formalism is derived likely generalize well to other forms of cultural innovation and contexts[63,119,120].

In our model, agents do not use the word until they have been exposed to it by a network neighbor at least once. Language change is better modeled in a usage-based rather than adopter-based framework (i.e., agents can use the word at each timestep rather than becoming and remaining an adopter one time)[18]. Accordingly, at each discrete timestep t, agent j decides whether they will use the word w with dynamic likelihood $p_{jwt} \in [0, 1]$, reflecting whether the word is salient to them[121]. This probability changes at each timestep[71,122], aggregating six pieces of information from agents' exposures to the new word: (i) **Attention Fading**: If agent j was previously exposed to the word but is not exposed at timestep t, their attention to the new word, and their likelihood of adoption, fades[121]. If agent j's network neighbor $i \in N(j)$ uses the word at timestep t (i.e., $i \in adopt(t)$), j updates their likelihood of using the word at the next timestep $p_{j,w,t+1}$. At this point, agent j's mental representations are determined by five main characteristics: (ii) **Novelty**: With greater exposure, a word's novelty wears off and its salience declines[123]. (iii) **Stickiness**: Some words are more likely to experience higher coinage and adoption because, for instance, they are related to topics of growing importance, used across a variety of semantic contexts, are associated with higher communicative need, or have notable linguistic properties[124–126]. (iv) **Relevance**: since speakers often use language to perform their own identity, agents may preferentially use words whose demographics more closely match their own[13,37]; (v) **Variety**: In addition to common identity, diverse exposure, from multiple people across multiple contexts, improves a word's salience and provides social affirmation for use of the word[118,127,128]; and (vi) **Relatability**: Since self-expression and social engagement are key motivators for use of social networking sites, input from agents with similar identity may weigh more heavily[61,76,129–131].

While many other factors may affect the diffusion of new words (cf. Supplementary Discussion), we do not include them in order to develop a parsimonious model that can be used to study specifically the effects of network and identity[132]. In particular, assumptions (iii)–(vi) are a fairly simple model of the effects of network and identity in the diffusion of lexical innovation. The network influences whether and to what extent an agent gets exposed to the word, using a linear-threshold-like adoption rule (assumption v) with a damping factor (assumption iii). Identity is modeled by allowing agents to both preferentially use words that match their own identity (assumption iv) and give higher weight to exposure from demographically similar network neighbors (assumption vi). Assumptions (i) and (ii) are optional to the study of network and identity and can be eliminated from the model when they do not apply (by removing Equation (1) or the η parameter from Equation (2)). For instance, these assumptions may not apply to more persistent innovations, whose adoption grows via an S-curve[58]. Since new words that appear in social media tend to be fads whose adoption peaks and fades away with time (Supplementary Fig. 8), we model the decay of attention theorized to underly this temporal behavior[133,134]. Without (i) and (ii), agents with a high probability of using the word would continue using it indefinitely. These assumptions allow the word to exit the lexicon and the cascade to stop.

Per Equation (1) and Equation (2), these six characteristics suggest that $p_{j,w,t+1}$ should be proportionate to: (i) **Attention Fading**: an exponential decay in attention[134], where agents retain fraction $r \in [0, 1]$ of their attention when not exposed to the word at time t:

$$p_{j,w,t+1} = r \cdot p_{jwt} \tag{1}$$

When agents are exposed at time t, $p_{j,w,t+1}$ is proportionate to (ii) **Novelty**: a cosine decaying function of the number of exposures j has had to the word η_{jwt}; (iii) **Stickiness:** the "stickiness" of the word S_w, which scales the probability of adoption; (iv) **Relevance:** the similarity between j's identity and their understanding of the word's identity, δ_{jw}; (v) **Variety**: the fraction of their network neighbors to have adopted the word at timestep t; and (vi) **Relatability**: this fraction is weighted by the similarity in their identity δ_{ij} and tie strength w_{ij}.

$$p_{j,w,t+1} = \delta_{jw} S_w \eta_{jwt} \frac{\sum\limits_{i \in N(j) \cap adopt(t)} w_{ij}\delta_{ij}}{\sum\limits_{k \in N(j)} w_{kj}\delta_{kj}} \tag{2}$$

In Equation (2), the network influences which words an agent has the opportunity to adopt and their likelihood of adopting those words by determining (1) the words an agent is exposed to and (2) the agents' level of exposure to the word. Identity is modeled in two ways: (1) agents preferentially use words that match their own identity (δ_{jw}), and (2) agents give higher weight to exposure from demographically similar network neighbors (δ_{ij}). In both mechanisms, new adopters would more likely be demographically similar and geographically proximal to existing adopters, producing geographic regions. Notably, agents may have a relatively high likelihood of adopting words if either the identity effect (word signals their identity) or the network effect (enough of their ego network is using the word) is sufficiently strong; in other words, an agent may have a reasonably high probability of adopting a word that doesn't signal their identity (which would make δ_{iw} low) if many of their friends are using it (which would make the last term in Equation (2) high).

Identity comparisons (δ_{jw}, δ_{ij}) are done component-wise, and then averaged using the weight vector v_w (section "Word identity"). Note that $p_{j,w,t+1}$ implicitly takes into account the value of $p_{j,w,t}$ by accounting for all exposures overall time. See Supplementary Methods 1.2.4 for the full set of model equations.

We stop the model once the growth in adoption slows to under 1% increase over ten timesteps. Since early timesteps have low adoption, uptake may fall below this threshold as the word is taking off; we reduce the frequency of such false-ends by running at least 100 timesteps after initialization before stopping the model.

Simulated counterfactuals

We directly assess the roles of network and identity in linguistic diffusion by evaluating the impact of omitting each of these sets of variables from the model. We simulate three counterfactual conditions to the full Network +Identity model described above:

- Network-only: eliminate agents performing identity by simulating the spread through just the weighted networks (δ_{ij}, $\delta_{jw} = 1$).
- Identity-only: shuffle the edges of the network. This configuration model-like procedure[135] preserves each agent's degree, allowing us to isolate the impact of eliminating homophily, the characteristic of the network most often hypothesized to drive regionalization, while also holding constant other network-geographic confounds like population and degree distributions.
- Null (Shuffled Network+No Identity): shuffled network without identity variables. This holds constant several variables (e.g., population size, degree distribution, model formulation), thus isolating the impact of structural factors other than network and identity.

Parameters and trials

We evaluate each model by examining its performance across 25 random trials on each of the 76 neologisms described in the section "New words and initial adopters" (1900 trials in total). In a sequence of three steps, non-empirical model parameters are tuned to the data and simulations are run at these parameters:

1. Parameters Q, r, and θ are tuned to the number of adoptions in a random 20% sample of words using a grid search. As described in Supplementary Methods 1.3, each parameter is assigned to the value that brings simulated usage (number of adoptions) closest to empirical usage; we do not maximize the study outcomes (e.g., Lee's L, likelihood of model pathways) and use a 20% sample instead of all words in order to avoid overfitting the model. The optimal values for these parameters are $Q = 0.75$, $r = 0.4$, and $\theta = 100$.
2. S_w is tuned separately for each word w, whereas in step #1, it is again fit to the number of adoptions using a grid search. As described in property (iii) of section "Diffusion", some words may be inherently more likely to be adopted than others. Therefore, each word takes on a different value of stickiness.
3. Five trials are run for each word w at the value of S_w from step #2.

Steps 2 and 3 are repeated five times, producing a total of 25 trials (five different stickiness values and five simulations at each value) per word, and a total of 1900 trials across all 76 words. This procedure is repeated on each of the four models from section "Simulated counterfactuals".

Model evaluation

We evaluate whether models match the empirical (i) spatial distribution of each word's usage and (ii) spatiotemporal pathways between pairs of counties.

First, we assess whether each model trial diffuses in a similar region as the word on Twitter. We compare the frequency of simulated and empirical adoptions per county using Lee's L, an extension of Pearson's R correlation that adjusts for the effects of spatial autocorrelation[136]. Based on Grieve et al. (2019)'s evaluation of this metric[107], the simulated and empirical regions are "very similar" if the correlation between the two spatial distributions is $L \geq 0.4$, "broadly similar" if $L \geq 0.13$, and "not similar" otherwise (see Supplementary Methods 1.5.2 for details).

Second, we compare the strength of empirical pathways against simulated pathways from the four models. The strength of the pathway between counties i and j is j's propensity to adopt the word after i does—measured via the zero-inflated correlation $\boldsymbol{\tau}$[137] between i's level of adoption at timestep t and j's adoption at $t+1$. We compare empirical to simulated pathways by calculating the Bayesian likelihood of the empirical pathway strengths $\boldsymbol{\tau}_E$ given the corresponding model pathway strengths $\hat{\boldsymbol{\tau}_{N+I}}$, $\hat{\boldsymbol{\tau}_N}$, or $\hat{\boldsymbol{\tau}}_I$. To validate this measure, we show that it reproduces ground truth pathways in simulated data. See Supplementary Methods 1.5.2 for more details on the metric and validation.

All reported differences are statistically significant at the level $\alpha = 0.05$, using a two-tailed bootstrap hypothesis test.

Hypotheses

Cultural artifacts like language often diffuse in well-known geographic regions. Our model formalizes two interacting mechanisms thought to generate this spatial heterogeneity: (1) network: edges tend to concentrate between demographically similar locales, meaning words may diffuse in regions well-connected by this network; and (2) identity: linguistic variants are selectively adopted in (and subsequently transmitted from) areas where speakers identify with their social signal (e.g., a word like "democrap" will likely get more use in a Republican-leaning area). Using this model, we test the roles of network and identity in diffusion.

In light of known urban/rural dynamics, our expectation is that network and identity are responsible for the spread of new words in different types of geographies. In particular, in diverse urban areas, we would expect new words to diffuse among dissimilar people via the network's weak ties. On the other hand, in more homogenous rural areas, we would expect these words to spread along strong ties with a shared identity. Consistent with this proposed mechanism, we hypothesize that:

H1. In the USA as a whole (across all urban and rural geographies), the Network+Identity model will outperform all other models, and the Null (Shuffled Network+No Identity) model will perform the worst.

H2. In different subsets of the country, network and identity may play more important roles. Specifically:

H2.1. Urban-Urban Diffusion: Transmission between two urban counties would be best approximated by the Network-only model.

H2.2. Rural-Rural Diffusion: Transmission between two rural (i.e., non-urban) counties would be best approximated by the Identity-only model.

H2.3. Urban-Rural Diffusion: Diffusion between an urban and a rural county (urban-to-rural or rural-to-urban) is best approximated by the Network+Identity model.

Note that, in testing these hypotheses, we do not penalize the Network+Identity model for added complexity. All models have the same number of free parameters that are tuned to the data. Moreover, our model predicts the spatial diffusion and pathways of a new word from first principles, unlike machine learning models that often learn these macroscopic patterns from the data. In a formal model, adding mechanisms that are unrelated to the process being simulated could result in a worse fit between the model's output and empirical data[138], so the Network+Identity model could have worse performance on a network- or identity-only process. Indeed, the Network+Identity model does not always outperform the Network- and Identity-only models: on average these counterfactuals better predict diffusion in urban and rural areas, respectively (see section "Network and identity play complementary, interacting roles"), and in 54% of the full-US simulations we ran, the Network- or Identity-only models had higher Lee's L correlation with the empirical geographical distribution (Network+Identity was best in 46% of trials, Network-only in 34% of trials, Identity-only in 20% of trials).

Testing the hypotheses

We run identically-seeded trials on all four models from section "Simulated counterfactuals" and track the number of adopters of each new word per county at each timestep. To test H1, we compare the performance of all four models on both metrics in section "Model evaluation".

To test H2, we classify each county as either urban or rural by adapting the US Office of Management and Budget's operationalization of the urbanized or metropolitan area vs. rural area dichotomy (see Supplementary Methods 2.8 for details). Then, using the measures from section 2.8, we calculate pathway weights and likelihoods between pairs of two urban counties (urban-urban), pairs of two rural counties (rural-rural), and between urban and rural counties (urban-rural, encompassing urban-to-rural or rural-to-urban).

In order to test whether network and identity play the hypothesized roles, we evaluate each model's ability to reproduce just urban-urban pathways, just rural-rural pathways, and just urban-rural pathways. Our hypotheses suggest that network or identity may better model urban and rural pathways alone rather than jointly. Our results are robust to removing location as a component of identity (Supplementary Methods 1.7.5), suggesting that our results are not influenced by explicitly modeling geographic identity.

To more directly test the proposed mechanism, we check whether the spread of new words across counties is more consistent with strong- or weak-tie diffusion. While our proposed mechanism is consistent with a purely empirical evaluation (network characteristics explain a higher fraction of the variation in Twitter's urban-urban pathway strength, while similarity in identity explains more in rural-rural empirical pathways (Supplementary Figs. 20, 21), these empirical characteristics likely have a nonlinear relationship with the strength of network- and identity-only pathways. Since we cannot empirically disentangle the network from identity, we use our Network-only model to assess whether pairs of counties are connected via a heavy network pathway (i.e., when the Network-only model pathway weight is high, suggesting diffusion occurs on the basis of network ties) and the Identity-only model to determine whether they are connected via a heavy identity pathway (i.e., when the Identity-only model pathway weight is high, suggesting diffusion occurs on the basis of shared identity).

Depending on the weight of the network- and identity-influenced pathways, diffusion between a pair of counties may tend to be driven by high levels of strong-tie diffusion (heavy network, heavy identity—or diffusion along network ties with shared identity); high levels of weak-tie diffusion (heavy network, light identity—or diffusion along diverse network ties); lower levels of strong-tie diffusion (light network, heavy identity); or low levels of weak-tie diffusion (light network, light identity). To check which of these mechanisms is most common in each type of geography, we use linear regression to correlate the strength of each empirical pathway ($\boldsymbol{\tau}_E$) to a three-way interaction between the strength of pathways in the Network- and Identity-only models ($\hat{\boldsymbol{\tau}_N}$, $\hat{\boldsymbol{\tau}}_I$) and the type of pathway (urban-urban, rural-rural, or urban-rural); see Supplementary Methods 1.5.3 for details.

Results

Network and identity better predict spatial properties jointly

Consistent with H1, we find that geographic properties of new words are best explained by the joint contributions of network and identity. Key properties of spatial diffusion include the frequency of adoption of innovation in different parts of the USA[23,67,139], as well as a new word's propensity to travel from one geographic area (e.g., counties) to another[23,67,139,140]. In both the physical and online worlds, where words are adopted carries signals about their cultural significance[21,141], while spread between pairs of counties acts like "pathways" along which, over time, variants diffuse into particular geographic regions[23,67,139].

Figure 1 shows the performance of all four models. Overall, the Network+Identity model best predicts a word's spatial diffusion. It is the only model

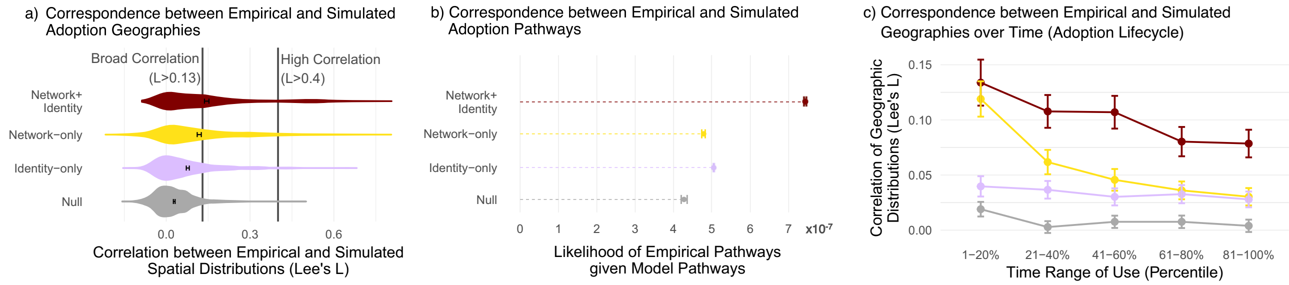

Fig. 1 | Model evaluation. The Network+Identity model best reproduces spatial diffusion on Twitter. **a** Shows the distribution of Lee's L correlations between simulated and empirical county maps, for all 1900 trials of each model; the black error bars show the 95% confidence interval for the mean correlation, and vertical lines are thresholds for "broadly" ($L > 0.13$) and "very similar" ($L > 0.4$) correlations. **b** Shows the likelihood of the pathways observed on Twitter given each of the simulations. **c** Shows the Lee's L correlation between the empirical and simulated geographic distributions over time; each point represents the Lee's L correlation between the geographies of adopters in each quintile (e.g., if there are 1000 empirical uses and 10,000 simulated of the word, the 20th–40th percentile of usage would be empirical uses #201–400 correlated with simulated uses #2001–4001). Error bars are 95% two-tailed bootstrap confidence intervals.

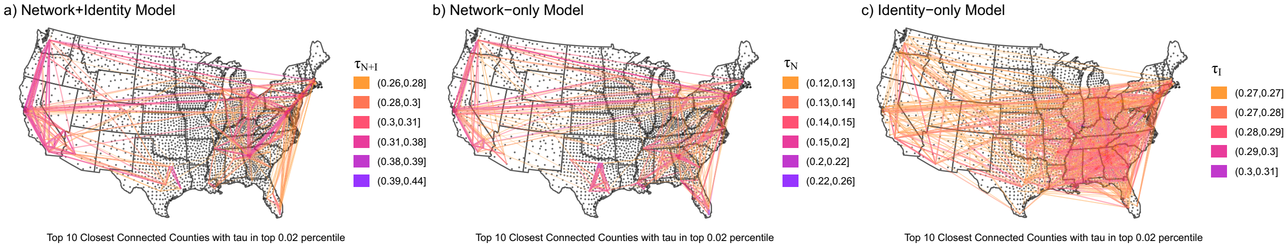

Fig. 2 | Model pathways. The Network+Identity model's pathways correspond to culturally significant regions. The maps depict the strongest pathways between pairs of counties in the **a** Network + Identity model, **b** Network-only model, and **c** Identity-only model. Pathways are shaded by their strength (purple is more strong, orange is less strong); if one county has more than ten pathways in this set, just the ten strongest pathways out of that county are pictured.

whose adoption regions are, on average, "broadly similar" to those on Twitter (mean(L) ≈ 0.15) (Fig. 1a), and the likelihood of the pathways observed on Twitter is more than 50% higher given the Network+Identity model's pathways than the other models' pathways (Fig. 1b). In turn, the Network- and Identity-only models far overperform the Null model on both metrics. These results suggest that spatial patterns of linguistic diffusion are the product of network and identity acting together. The Network- and Identity-only models have diminished capacity to predict geographic distributions of lexical innovation, potentially attributable to the failure to effectively reproduce the spatiotemporal mechanisms underlying cultural diffusion. Additionally, both network and identity account for some key diffusion mechanism that is not explained solely by the structural factors in the Null model (e.g., population density, degree distributions, and model formulation).

Note that, for the sake of interpretability, our model is very simple (e.g., built on first principles, one parameter S_w tuned, and initialized with only the word's first ten adopters), and a more complex model (e.g., better trained to the data) would likely have even higher performance. However, in spite of this, the Network+Identity model is able to capture many key spatial properties. Nearly 40% of Network+Identity simulations are at least "broadly similar," and 12% of simulations are "very similar" to the corresponding empirical distribution (Fig. 1a). The Network+Identity model's Lee's L distribution roughly matches the distribution Grieve et al. (2019) found for regional lexical variation on Twitter, suggesting that the Network +Identity model reproduces "the same basic underlying regional patterns" found on Twitter[107]. Compared to other models, the Network+Identity model was especially likely to simulate geographic distributions that are "very similar" to the corresponding empirical distribution (12.3 vs. 6.8 vs. 3.7%). These "very similar" distributions tended to occur among words whose adopters are highly localized (average Moran's I of 0.84 among very similar vs. 0.66 among others) and where the Network- or Identity-only models tend to have a "very similar" distribution (34 and 20%, respectively —in these cases, the Network+Identity model almost always improves upon the performance of the Network- and Identity-only counterfactuals). These results suggest that network and identity are particularly effective at modeling the *localization* of language.

Figure 2 shows the strongest spatiotemporal pathways between pairs of counties in each model. Visually, the Network+Identity model's strongest pathways correspond to well-known cultural regions (Fig. 2a). Some pathways extend from the mid-Atlantic into the South, where African American Language is most spoken[94]; from Atlanta to other urban hubs, along pathways defined by the Great Migrations[94]; along and between both coasts, which are politically, linguistically, and racially distinctive from the middle of the country[14,100]; within the economically significant Dallas-Austin-Houston "Texas triangle"[142]; and between this Texas region and the West Coast[143]. These pathways likely capture the complementary effects of network and identity. The Network-only model does not capture the Great Migration or Texas-West Coast pathways (Fig. 2b), while the Identity-only model only produces just these two sets of pathways but none of the others (Fig. 2c). These results suggest that network and identity reproduce the spread of words on Twitter via distinct, socially significant pathways of diffusion. Our model appears to reproduce the mechanisms that give rise to several well-studied cultural regions.

Notably, the Network+Identity model is best able to reproduce spatial distributions over the entire lifecycle of a word's adoption. Figure 1c shows how the correlation between the empirical and simulated geographic distributions changes over time. Early adoption is well-simulated by the network alone, but later adoption is better simulated by network and identity together as the Network-only model's performance rapidly deteriorates over time. The Identity-only and Null models perform poorly at all times. These results are consistent with H2, since theory suggests that early adoption occurs in urban areas (which H2 suggests would be best modeled by network alone) and later adoption is urban-to-rural or rural-to-rural (best modeled by network+identity or identity alone, per H2)[25]. We will more directly test H2 in the next section.

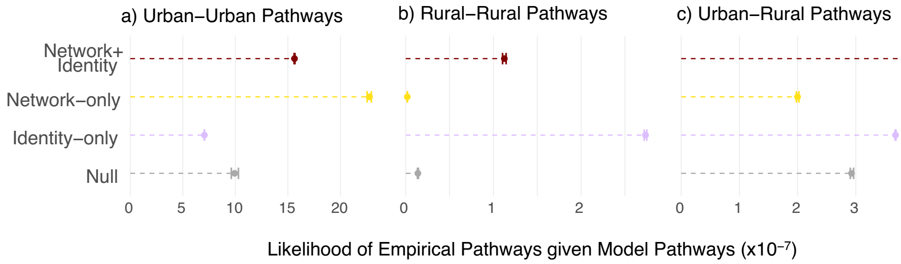

Fig. 3 | Urban/rural evaluation. Based on the likelihood of the pathways observed on Twitter given each of the simulations: **a**) The Network-only model best matches pathways containing an urban county; **b**) The Identity-only model best matches pathways among rural counties; and **c**) the Network+Identity model best matches pathways connecting an urban county to a rural county. Error bars are 95% two-tailed bootstrap confidence intervals.

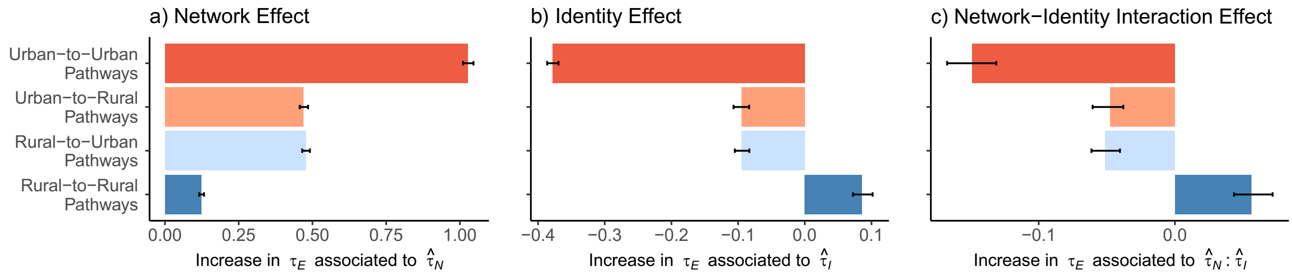

Fig. 4 | Urban/rural mechanisms. Based on standardized coefficients from a linear regression predicting empirical pathway strength ($\boldsymbol{\tau}_E$) from a three-way interaction between the strength of the pathways in the Network- and Identity-only models ($\hat{\boldsymbol{\tau}}_N$, $\hat{\boldsymbol{\tau}}_I$) and the type of pathway (urban vs. rural county): **a** The strength of the Network-only model's pathways have the largest effect on the strength of the urban-urban empirical pathways and are positively associated with all pathways; **b** Conversely, identity pathways have the largest effect on the strength of rural-rural pathways and is negatively associated with urban pathways; and **c** Urban heavy network pathways are weakened by heavy identity pathways—and conversely, rural-rural heavy identity pathways are strengthened by heavy network pathways. Error bars are 95% two-tailed bootstrap confidence intervals.

Network and identity play complementary, interacting roles

Next, we show that network- and identity-influenced pathways between counties play distinct roles in the spread of innovation. As expected, pathway strengths in the Network- and Identity-only models are strongly correlated (Pearson's $R = 0.78$, Spearman's $\rho = 0.81$), since edges in the network often form between demographically similar individuals[49] (see Supplementary Methods 1.6.4 for details). Nonetheless, the Network- and Identity-only pathways exhibit important differences, and our hypothesis is that spatial diffusion in the USA consists of two interacting mechanisms: The adoption of innovation among urban counties tends to happen via weak-tie diffusion—because for multiple reasons, potentially including structural factors like the preponderance of weak and demographically dissimilar ties or behavioral factor like preferences for diverse input[144,145], urban diffusion may tend to occur when demographically dissimilar speakers are exposed to words that have not yet entered their social circle. Among rural counties, on the other hand, we expect new cultural artifacts to spread via strong-tie diffusion; speakers are largely connected to demographically-like individuals via strong ties, and adopt words that signal an identity that both parties share. Evidence from social networking sites suggests that urban vs. rural heterogeneity persists online[146], suggesting that this mechanism is testable in our setting.

We find that, although network- or identity-only models may show promising results in one type of geography, these same models will not work in all subsets of the USA. Figure 3 quantifies the efficacy of network and identity in urban and rural diffusion, while Fig. 4 shows the associations between the empirical pathway strength and the Network- and Identity-only strengths ($\hat{\boldsymbol{\tau}}_N$, $\hat{\boldsymbol{\tau}}_I$) in these different geographies. We find that H2.1) the Network-only model best explains the strength of urban-urban pathways; H2.2) the Identity-only model most closely approximates empirical rural-rural pathways; and H2.3) the strength of urban-rural pathways is best captured by the joint Network+Identity model. To elaborate:

H2.1: Weak-tie diffusion along urban-urban pathways. Empirical pathways are heaviest when there is a heavy network and light identity pathway (high levels of weak-tie diffusion) and lightest when both network and identity pathways are heavy (high levels of strong-tie diffusion) (Fig. 4, dark orange bars). In other words, diffusion between pairs of urban counties tends to occur via weak-tie diffusion—spread between dissimilar network neighbors connected by low-weight ties[76]. This is consistent with Fig. 3a, where the Network-only model best reproduces the weak-tie diffusion mechanism in urban-urban pathways; conversely, the Identity-only and Network+Identity models perform worse in urban-urban pathways, amplifying strong-tie diffusion among demographically similar ties.

H2.2: Strong-tie diffusion along rural-rural pathways. Empirical rural-rural pathways tend to be heavier when both network and identity pathways are heavy (high levels of strong-tie diffusion), and lightest when both network and identity pathways are light (low levels of weak-tie diffusion) (Fig. 4, dark blue bars). This suggests that transmission between two rural counties tends to occur via strong-tie diffusion. This is consistent with Fig. 3b, where the Identity-only model best reproduces strong-tie diffusion among rural-rural pathways, increasing spread among only counties with relevant shared identities; conversely, the Network-only and Network+Identity models underperform by inflating levels of diffusion among strongly connected individuals who lack a relevant shared identity. For example, if two strongly tied speakers share a political but not linguistic identity, the identity-only model would differentiate between words signaling politics and language, but the network-only model would not.

H2.3: Network and identity required for diffusion between urban and rural areas. Finally, pathways between an urban and a rural county (urban-to-rural or rural-to-urban) tend to fall in between urban-urban and rural-rural pathways—relying more on identity than urban-urban pathways and more on the network than the rural-rural pathways (Fig. 4, light orange/blue bars). As such, the Network+Identity model, which includes both factors, best predicts these pathway strengths in Fig. 3c. These results suggest that network and identity may both be involved in a word spreading between urban and rural counties—for instance, a network- or identity-only model of diffusion may not explain urban-rural diffusion well, because words may travel from an urban center to a more sparsely populated rural area via both weak ties (diverse connections, bridging different geographic regions) and strong ties (geographically distal but socially proximal connections, perhaps remnants of migrations or other contact[27]).

Although differences in cultural diffusion between urban and rural areas have been well-documented[24–27,29–31], few prior studies could explain how these differences came to be. We offer a well-reasoned proposal as to how network and identity produce these patterns. Specifically, these two social structures take on complementary, interacting functions: identity pathways drive transmission among rural counties via strong-tie diffusion, while network pathways dominate urban-urban spread via weak-tie diffusion. The interaction of network, identity, and type of pathway explains a high fraction (almost 70%) of the variance in empirical pathway strength. Empirical pathways, then, are well-explained by our proposed mechanism, since most of the variance in the strength of pathways can be explained by urban/rural differences in weak- and strong-tie diffusion.

Furthermore, as shown in Supplementary Methods 1.6.5, urban/rural dynamics are only partially explained by distributions of network and identity. The Network+Identity model was able to replicate most of the empirical urban/rural associations with network and identity (Supplementary Fig. 17), so empirical distributions of demographics and network ties likely drive many urban/rural dynamics. However, unlike empirical pathways, the Network+Identity model's urban-urban pathways tend to be *heavier* in the presence of heavy identity pathways, since agents in the model select variants on the basis of shared identity. These results suggest that urban-urban weak-tie diffusion requires some mechanism not captured in our model, such as urban speakers seeking diversity or being less attentive to identity than rural speakers when selecting variants[144,145].

Finally, contrary to prior theories[24,25,147], properties like population size and the number of incoming and outgoing ties were insufficient to reproduce urban/rural differences. The Null model, which has the same population and degree distribution, underperformed the Network+Identity model in all types of pathways. However, notably, the Null model predicts urban-urban pathway strengths better than identity alone and rural-rural pathway strengths better than network alone, suggesting that population distributions and other structural properties may be a better predictor of diffusion than network or identity alone in some geographies, and underscoring the fact that network and identity facilitate complementary mechanisms of diffusion that are each necessary in different parts of USA.

Overall, both network and identity are required to explain the adoption of innovation: omitting either one entails not only poorer prediction of spatial properties, but also losing a key determinant of diffusion. Because of these interacting mechanisms, innovation may be adopted less selectively in urban areas, where populations are more diverse and more likely connected by weak ties, and words may diffuse along strong ties in the more homogeneous rural areas if they signal a shared identity.

Discussion

We demonstrate that many existing models of cultural diffusion are missing a key dynamic in the adoption of innovation: models that consider identity alone ignore weak-tie diffusion between an urban resident and their diverse contacts; while models that use network alone are unable to consider shared identity and, as a result, likely dilute the diffusion of local variants to and from rural areas. One direct consequence, as demonstrated by the simulated counterfactuals, is a loss of accuracy in reproducing spatial distributions and spatiotemporal pathways of diffusion. Moreover, the absence of either network or identity also hamstrings a model's ability to reproduce key macroscopic dynamics like urban-rural diffusion that are likely the product of both strong-tie and weak-tie spread.

We also propose and test a mechanism through which words diffuse between and among urban or rural areas. Through this framework, we see that the adoption of cultural innovation is the product of complementary, interacting roles of network and identity. These ideas build on a rich literature on the mechanisms of spatial diffusion[148–150] and have powerful theoretic implications across disciplines. In the subfield of variationist sociolinguistics, our proposed mechanism for diffusion draws a link between identity- and network-based explanations of language change[35]: showing how strong- and weak-tie theory require information about network and identity to work together. In network theory, this idea suggests how strong ties may influence diffusion when reinforced by node characteristics like identity[47], and integrate Granovetter's theories on tie strength[76] with cultural theory about the role of urban centers and rural peripheries in diffusion[25,27]. Moreover, in cultural geography, our analysis provides a key contribution to theory: since urban vs. rural differences are emergent properties of our model's minimal assumptions, urban/rural variation may not be the result of the factors to which it is commonly attributed (e.g., population size and edge distribution). Instead, people perform their spatially-correlated identities by choosing among variants that diffuse through homophilous networks; the differences in network topology and demographic distributions in urban and rural populations, then, may create the observed differences in adoption. Importantly, our results suggest that, urban and rural populations both contribute differently to the diffusion of cultural innovation, rather than there being one dominating culture online. The geographic regions found with our data also highlight that despite the ease of widespread dissemination of cultural artifacts in online settings which could lead to more universally-shared behaviors, pre-Internet geographic distinctions in culture still persist.

Although our hypotheses were tested on lexical diffusion in the USA, the results may apply to the spread of many other types of cultural innovation (e.g., music, beliefs) in a single country or even globally. Linguistic variants often serve as proxies for cultural variables, since their adoption tends to reflect broader societal shifts[10–14,17]. Although many of our assumptions about spatial patterns may not apply in every part of the world (e.g., in places that are less diverse or spatially segregated), the model may also apply to other countries or even international contexts where networks and identities are geographically correlated[146]. In these cases, however, it would be important to adapt how one estimates network and identity: e.g., the network may be better estimated using platforms other than Twitter or even surveys, and salient identities may not be demographic. Additionally, the type of geographic patterns we found relied on there being one type of geography where weak-tie (diverse) diffusion was more common and other where strong-tie (shared identity) diffusion was more common and our results are unlikely to generalize to areas where this is not the case. This sort of mechanism, combining strong and weak-tie diffusion, has been hypothesized in cross-country diffusion of business models[151], and could be applicable to other forms of innovation as well.

Moreover, the assumptions of our model are sufficiently general to apply to the adoption of many social or cultural artifacts. However, since our model assumes a non-zero probability of adoption from the start, it likely would apply only to forms of innovation where the barriers to adoption are low enough for the effects of network and identity to be salient (e.g., not something like technological innovation where functional needs and accessibility are factors). We might also expect the Network-only model to perform best when weak-tie diffusion is the main mechanism (e.g., job information[76]) and the Identity-only model to perform better when innovation spreads mainly through strong-tie diffusion (e.g., health behaviors, activism[152,153]). Importantly, our conclusions about the importance of network and identity, and the mechanisms we have identified for their interaction, may have applicability across a range of social science disciplines—and future work can use the agent-based model developed in this paper to test whether these findings generalize to other cultural domains.

In order to make more accurate predictions about how innovation diffuses, we call on researchers across disciplines to incorporate both network and identity in their (conceptual or computational) models of diffusion. Scholars can develop and test theory about the ways in which other place-based characteristics (e.g., diffusion into specific cultural regions) emerge from network and identity. Our model has many limitations (detailed in Supplementary Discussion), including that our only data source was a 10% Twitter sample, our operationalization of network and identity, and several simplifying assumptions in the model. Nevertheless, our work offers one methodology, combining agent-based simulations with large-scale social datasets, through which researchers may create a joint network/identity model and use it to test hypotheses about mechanisms underlying cultural diffusion.

Data availability

The datasets pertaining to the new words identified in this study (word list, initial adopters, identities signaled, day/county-level spatial timeseries) are available on Github: https://github.com/aparna-ananth/network-identity-abm. The Twitter network (edgelist) and users (registry) that support the findings of this study are taken from our university's Twitter Decahose, but restrictions apply to the availability of these data, which were used under licence for the current study and so are not publicly available.

Code availability

All code for the models and analysis in this study are available on Github: https://github.com/aparna-ananth/network-identity-abm.

Received: 27 October 2023; Accepted: 13 June 2024;
Published online: 02 September 2024

References

1. Sauer, C. O. *Agricultural Origins and Dispersals* (The American Geographical Society, 1952).
2. Gould, P. R. *Spatial Diffusion, Resource Paper No. 4*. Report No. ED120029C https://eric.ed.gov/?id=ED120029 (Association of American Geographers,1969).
3. Kong, L. Geography and religion: trends and prospects. *Prog. Hum. Geogr.* **14**, 355–371 (1990).
4. Land, K. C., Deane, G. & Blau, J. R. Religious pluralism and church membership: a spatial diffusion model. *Am. Sociol. Rev.* **56**, 237–249 (1991).
5. Kellogg, A. E. *Spatial Diffusion of Popular Music via Radio in the United States*. Ph.D. thesis, Michigan State University (1986).
6. Nash, P. H. & Carney, G. O. The seven themes of music geography. *Can. Geogr.* **40**, 69–74 (1996).
7. Kamath, K. Y., Caverlee, J., Lee, K. & Cheng, Z. Spatio-temporal dynamics of online memes: a study of geo-tagged tweets. In *Proc. 22nd International Conference on World Wide Web* 667–678 (Association for Computing Machinery, 2013).
8. Dang, L., Chen, Z., Lee, J., Tsou, M.-H. & Ye, X. Simulating the spatial diffusion of memes on social media networks. *Int. J. Geogr. Inf. Sci.* **33**, 1545–1568 (2019).
9. Woodard, C. *American Nations: A History of the Eleven Rival Regional Cultures of North America* (Penguin, 2011).
10. Jackson, P. *Maps of Meaning* (Routledge, 2012).
11. Chambers, J. K. & Trudgill, P. Dialectology (Cambridge Univ. Press, 1998).
12. Grieve, J. *Regional Variation in Written American English* (Cambridge Univ. Press, 2016).
13. Labov, W. The social motivation of a sound change. *Word* **19**, 273–309 (1963).
14. Labov, W. *Dialect Diversity in America: The Politics of Language Change* (University of Virginia Press, 2012).
15. Bail, C. A. The cultural environment: measuring culture with big data. *Theory Soc.* **43**, 465–482 (2014).
16. Anderson, J. *The Places and Traces of Language* (Routledge, 2021).
17. Kramsch, C. *Language and Culture* (Oxford Univ. Press, 1998).
18. Beckner, C. et al. Language is a complex adaptive system. *Lang. Learn.* **11**, 1–26 (2009).
19. Kramsch, C. Language and culture. *AILA Rev.* **27**, 30–55 (2014).
20. Hagerstrand, T. *Innovation Diffusion as a Spatial Process* (University of Chicago Press, 1967).
21. Trudgill, P. Linguistic change and diffusion: description and explanation in sociolinguistic dialect geography. *Lang. Soc.* **3**, 215–246 (1974).
22. Trudgill, P. *Sociolinguistic Variation and Change* (Edinburgh Univ. Press, 2002).
23. Labov, W., Ash, S. & Boberg, C. *The Atlas of North American English: Phonetics, Phonology and Sound Change* (Walter de Gruyter, 2008).
24. Fischer, C. S. Urban-to-rural diffusion of opinions in contemporary america. *Am. J. Sociol.* **84**, 151–159 (1978).
25. Labov, W. in *Social Dialectology: In Honour of Peter Trudgill* (eds Britain, D., Cheshire, J. & Trudgill, P.) Ch. 2, 9–22 https://benjamins.com/catalog/impact.16.03lab (John Benjamins Pub., 2003).
26. Brunstad, E., Røyneland, U. & Opsahl, T. *Hip Hop, Ethnicity and Linguistic Practice in Rural and Urban* (Continuum International Publishing Group, 2010).
27. Stewart Jr, C. T. The urban-rural dichotomy: concepts and uses. *Am. J. Sociol.* **64**, 152–158 (1958).
28. Fagyal, Z., Swarup, S., Escobar, A. M., Gasser, L. & Lakkaraju, K. Centers and peripheries: network roles in language change. *Lingua* **120**, 2061–2079 (2010).
29. Agergaard, J., Fole, N. & Gough, K. *Rural-Urban Dynamics* (Routledge, 2015).
30. Trudgill, P. et al. in *Language and Space: Theories and Methods* (eds. Peter A. & Jürgen E. S.) Ch. 18 https://www.degruyter.com/document/doi/10.1515/9783110220278.fm/pdf (De Gruyter Mouton, 2010).
31. Lengyel, B., Bokányi, E., Di Clemente, R., Kertész, J. & González, M. C. The role of geography in the complex diffusion of innovations. *Sci. Rep.* **10**, 1–11 (2020).
32. Lengyel, B., Varga, A., Ságvári, B., Jakobi, Á. & Kertész, J. Geographies of an online social network. *PLoS ONE* **10**, e0137248 (2015).
33. Lengyel, B. & Jakobi, Á. Online social networks, location, and the dual effect of distance from the centre. *Tijdschr. Econ. Soc. Geogr.* **107**, 298–315 (2016).
34. Bokányi, E., Novák, M., Jakobi, Á. & Lengyel, B. Urban hierarchy and spatial diffusion over the innovation life cycle. *R. Soc. Open Sci.* **9**, 211038 (2022).
35. Labov, W. Transmission and diffusion. *Language* **83**, 344–387 (2007).
36. Sturtevant, E. H. *An Introduction to Linguistic Science* (Yale Univ. Press, 1947).
37. Eckert, P. Variation and the indexical field 1. *J. Socioling.* **12**, 453–476 (2008).
38. Eckert, P. The whole woman: sex and gender differences in variation. *Lang. Var. Change* **1**, 245–267 (1989).
39. Labov, W. *The Social Stratification of English in New York City* (Cambridge Univ. Press, 2006).
40. Goel, R. et al. in *International Conference on Social Informatics* (Springer, 2016).
41. Schwartz, R. & Halegoua, G. R. The spatial self: location-based identity performance on social media. *New Media Soc.* **17**, 1643–1660 (2015).
42. Bloomfield, L. *Language* (Univ. Chicago Press, 1933).
43. Milroy, L. *Language and Social Networks* (Wiley-Blackwell, 1987).
44. Aral, S., Muchnik, L. & Sundararajan, A. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proc. Natl Acad. Sci. USA* **106**, 21544–21549 (2009).
45. Jackson, M. O. & López-Pintado, D. Diffusion and contagion in networks with heterogeneous agents and homophily. *Netw. Sci.* **1**, 49–67 (2013).
46. Toole, J. L., Cha, M. & González, M. C. Modeling the adoption of innovations in the presence of geographic and media influences. *PLoS ONE* **7**, e29528 (2012).
47. Milroy, L. & Milroy, J. Social network and social class: toward an integrated sociolinguistic model. *Lang. Soc.* **21**, 1–26 (1992).

48. Zhu, J. & Jurgens, D. The structure of online social networks modulates the rate of lexical change. In *Proc. North American Meeting of the Association for Computational Linguistics (NAACL)* (Association for Computational Linguistics, 2021).
49. McPherson, M., Smith-Lovin, L. & Cook, J. M. Birds of a feather: homophily in social networks. *Ann. Rev. Sociol.* **27**, 415–444 (2001).
50. Adamic, L. A. & Adar, E. Friends and neighbors on the web. *Social Netw.* **25**, 211–230 (2003).
51. Lizardo, O. How cultural tastes shape personal networks. *Am. Sociol. Rev.* **71**, 778–807 (2006).
52. Takhteyev, Y., Gruzd, A. & Wellman, B. Geography of twitter networks. *Soc. Netw.* **34**, 73–81 (2012).
53. Jackson, P. Rematerializing social and cultural geography. *Soc. Cult. Geogr.* **1**, 9–14 (2000).
54. Newman, M. E., Barabási, A.-L. E. & Watts, D. J. *The Structure and Dynamics of Networks* (Princeton Univ. Press, 2006).
55. Weinreich, U., Labov, W. & Herzog, M. I. in *Directions for Historical Linguistics* (eds Lehmann, W. P. & Malkiel, Y.) (University of Texas Press, 1968).
56. Brown, L. A. *Innovation Diffusion; A New Perspective* (Methuen, 1981).
57. Palloni, A.Diffusion in sociological analysis, 67–114 (National Academies Press Washington, DC, 2001).
58. Blythe, R. A. & Croft, W. S-curves and the mechanisms of propagation in language change. *Language* **2**, 269–304 (2012).
59. Bakshy, E., Rosenn, I., Marlow, C. & Adamic, L. The role of social networks in information diffusion. In *Proc. 21st international conference on World Wide Web* (Association for Computing Machinery, 2012).
60. Marshall, B. D. & Galea, S. Formalizing the role of agent-based modeling in causal inference and epidemiology. *Am. J. Epidemiol.* **181**, 92–99 (2015).
61. Valente, T. W. Social network thresholds in the diffusion of innovations. *Soc. Netw.* **18**, 69–89 (1996).
62. Valente, T. W. Network models of the diffusion of innovations. *J. Market.* **60**, 134 (1996).
63. Hruschka, D. J. et al. Building social cognitive models of language change. *Trends Cog. Sci.* **13**, 464–469 (2009).
64. Albright, A. 'The Dynamic Lexicon', In *the Oxford Handbook of Laboratory Phonology* (eds Abigail C. Cohn, C. F., & Marie K. H), https://doi.org/10.1093/oxfordhb/9780199575039.013.0008 (Oxford Academic, 2012).
65. Crystal, D. *Internet Linguistics: A Student Guide* (Routledge, 2011).
66. Kerremans, D., Stegmayr, S. & Schmid, H.-J. *Current Methods in Historical Semantics* (De Gruyter Mouton, 2012).
67. Eisenstein, J., O'Connor, B., Smith, N. A. & Xing, E. P. Mapping the geographical diffusion of new words. In *Proc. NIPS Workshop on Social Network and Social Media Analysis: Methods, Models and Applications* (Citeseer, 2012).
68. Miller, D. G. *English Lexicogenesis* (Oxford Univ. Press, 2014).
69. Grieve, J., Nini, A. & Guo, D. Mapping lexical innovation on american social media. *J. Engl. Linguist.* **46**, 293–319 (2018).
70. Rogers, E. M. *Diffusion of Innovations* (Simon and Schuster, 2010).
71. Agha, A. The social life of cultural value. *Lang. Commun.* **23**, 231–273 (2003).
72. Aral, S. & Dhillon, P. S. Social influence maximization under empirical influence models. *Nat. Hum. Behav.* **2**, 375–382 (2018).
73. DiMaggio, P. *Cultural Networks* (Sage, 2011).
74. Breiger, R. L. & Puetz, K. in *International Encyclopedia of Social and Behavioral Sciences* (Citeseer, 2015).
75. DiMaggio, P. & Cohen, J. in *The Economic Sociology of Capitalism* (Princeton Univ. Press, 2021).
76. Granovetter, M. The strength of weak ties. *Am. J. Sociol.* **78**, 1360 – 1380 (1973).
77. Gupte, M. & Eliassi-Rad, T. Measuring tie strength in implicit social networks. In *Proc. 4th Annual ACM Web Science Conference* 109–118 (Association for Computing Machinery, 2012).
78. Huberman, B., Romero, D. M. & Wu, F. Social networks that matter: Twitter under the microscope. *First Monday* **14** (2008).
79. Romero, D., Tan, C. & Ugander, J. On the interplay between social and topical structure. In *Proceedings of the International AAAI Conference on Web and Social Media* **7**, 516–525 https://doi.org/10.1609/icwsm.v7i1.14411(2013).
80. Zhu, Y.-X. et al. Influence of reciprocal links in social networks. *PLoS ONE* **9**, e103007 (2014).
81. Leskovec, J., Huttenlocher, D. & Kleinberg, J. Signed networks in social media. In *Proc. SIGCHI conference on human factors in computing systems* 1361–1370 (ACM, 2010).
82. He, X., Du, H., Feldman, M. W. & Li, G. Information diffusion in signed networks. *PLoS ONE* **14**, e0224177 (2019).
83. Gilbert, E., Karahalios, K. & Sandvig, C. The network in the garden: an empirical analysis of social media in rural life. In *Proc. SIGCHI Conference on Human Factors in Computing Systems* 1603–1612 (ACM, 2008).
84. Bailey, M., Cao, R., Kuchler, T., Stroebel, J. & Wong, A. Social connectedness: measurement, determinants, and effects. *J. Econ. Perspect.* **32**, 259–280 (2018).
85. Friedman, J. Culture, identity, and world process. *Review (Fernand Braudel Center)* **12**, 51–69 (1989).
86. Côté, J. E. Sociological perspectives on identity formation: the culture–identity link and identity capital. *J. Adolesc.* **19**, 417–428 (1996).
87. Jones, S. *Virtual Culture: Identity and Communication in Cybersocie*ty (Sage, 1997).
88. Eckert, P. Three waves of variation study: the emergence of meaning in the study of sociolinguistic variation. *Ann. Rev Anthropol.* **41**, 87–100 (2012).
89. Carver, C. M. *American Regional Dialects: A Word Geograp*hy (Univ. Michigan Press, 1987).
90. Eisenstein, J., O'Connor, B., Smith, N. A. & Xing, E. P. A latent variable model for geographic lexical variation. In *Proc. 2010 Conference on Empirical Methods in Natural Language Processing* 1277–1287 (Association for Computational Linguistics, 2010).
91. Rickford, J. R. *African American Vernacular English: Features, Evolution, Educational Implications* (Wiley, 1999).
92. Fought, C. *Chicano English in Context* (Springer, 2002).
93. Stewart, I. Now we stronger than ever: African-american english syntax in twitter. In *Proc. Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics* 31–37 (Association for Computational Linguistics, 2014).
94. Jones, T. Toward a description of african american vernacular english dialect regions using "black twitter". *Am. Speech* **90**, 403–440 (2015).
95. Labov, W. The intersection of sex and social class in the course of linguistic change. *Lang. Var. Change* **2**, 205–254 (1990).
96. Abitbol, J. L., Karsai, M., Magué, J.-P., Chevrot, J.-P. & Fleury, E. Socioeconomic dependencies of linguistic patterns in twitter: a multivariate analysis. In *Proc. 2018 World Wide Web Conference* 1125–1134 (2018).
97. Haugen, E. The analysis of linguistic borrowing. *Language* **26**, 210–231 (1950).
98. Lo, A. Codeswitching, speech community membership, and the construction of ethnic identity. *J. Socioling.* **3**, 461–479 (1999).
99. Stewart, I., Pinter, Y. & Eisenstein, J. Si o no, que penses? catalonian independence and linguistic identity on social media. In *Proc. 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)* 136–141 (Association for Computational Linguistics, 2018).
100. Sylwester, K. & Purver, M. Twitter language use reflects psychological differences between democrats and republicans. *PLoS ONE* **10**, e0137422 (2015).

101. Compton, R., Jurgens, D. & Allen, D. Geotagging one hundred million twitter accounts with total variation minimization. In *2014 IEEE International Conference on Big Data (big data)* 393–401 (IEEE, 2014).
102. Johnson, I., McMahon, C., Schöning, J. & Hecht, B. The effect of population and" structural" biases on social media-based algorithms: a case study in geolocation inference across the urban-rural spectrum. In *Proceedings of the 2017 CHI conference on Human Factors in Computing Systems*, 1167–1178 (ACM, 2017).
103. Buolamwini, J. & Gebru, T. Gender shades: intersectional accuracy disparities in commercial gender classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* **81**, 77–91 https://proceedings.mlr.press/v81/buolamwini18a.html (2018).
104. Keyes, O. The misgendering machines: trans/hci implications of automatic gender recognition. In *Proc. ACM on Human-Computer Interaction* 1–22 (ACM, 2018).
105. Xiong, C., Hu, S., Yang, M., Luo, W. & Zhang, L. Mobile device data reveal the dynamics in a positive relationship between human mobility and covid-19 infections. *Proc. Natl Acad. Sci. USA* **117**, 27087–27089 (2020).
106. Wrigley-Field, E. Us racial inequality may be as deadly as covid-19. *Proc. Natl Acad. Sci.* **117**, 21854–21856 (2020).
107. Grieve, J., Montgomery, C., Nini, A., Murakami, A. & Guo, D. Mapping lexical dialect variation in british english using twitter. *Front. Artif. Intell.* **2**, 11 (2019).
108. States, U. Census tracts and block numbering areas Chap. 10 (U.S. Dept. of Commerce, Economics and Statistics Administration, Bureau of the Census, 1994).
109. Powell, R., Clark, J. & Dube, M. *Assessing the Causes of District Homogeneity in US House Elections*. Report No. 2017-22 MIT Political Science Department Research Paper (2017).
110. Agha, A. Voice, footing, enregisterment. *J. Ling. Anthropol.* **15**, 38–59 (2005).
111. Goffman, E. et al. *The Presentation of Self in Everyday Life* Vol. 21 (Harmondsworth, 1978).
112. Oring, E. The arts, artifacts, and artifices of identity. *J. Am. Folk.* **107**, 211–233 (1994).
113. Wagner, A. M. *Gettin'weird Together: The Performance of Identity and Community through Cultural Artifacts of Electronic Dance Music Culture*. MSc thesis, Illinois State Univ. (2014).
114. Eckert, P. *Language Variation as Social Practice: The Linguistic Construction of Identity in Belten High* (Wiley, 2000).
115. Blommaert, J. in *Translinguistics* (Routledge, 2019).
116. Ilbury, C. "Sassy queens": stylistic orthographic variation in twitter and the enregisterment of aave. *J. Socioling.* **24**, 245–264 (2020).
117. Ellis, N. C. Frequency effects in language processing: a review with implications for theories of implicit and explicit language acquisition. *Stud. Second Lang. Acquis.* **24**, 143–188 (2002).
118. Centola, D. & Macy, M. Complex contagions and the weakness of long ties. *Am. J. Sociol.* **113**, 702–734 (2007).
119. DiMaggio, P. Culture and cognition. *Ann. Rev. Sociol.* **23**, 263–287 (1997).
120. DiMaggio, P. & Markus, H. R. Culture and social psychology: converging perspectives. *Soc. Psychol. Q.* **73**, 347–352 (2010).
121. Ellis, N. C. Essentials of a theory of language cognition. *Mod. Lang. J.* **103**, 39–60 (2019).
122. Kirby, S., Griffiths, T. & Smith, K. Iterated learning and the evolution of language. *Curr. Opin. Neurobiol.* **28**, 108–114 (2014).
123. Ellis, N. C. Salience in Language Use, Learning, and Change. In *The Changing English Language* Ch. 4 (eds. Hundt, M., Mollin, S. & Pfenninger, S) https://www.cambridge.org/core/books/abs/changing-englishlanguage/contents/250EDDA6783F1EF767CCBEB7B8410D5D (Cambridge Univ. Press, 2017).
124. Schmid, H.-J. & Günther, F. Toward a unified socio-cognitive framework for salience in language. *Front. Psychol.* **7**, 1110 (2016).
125. Stewart, I. & Eisenstein, J. Making “fetch” happen: the influence of social and linguistic context on nonstandard word growth and decline. In *Proc. 2018 Conference on Empirical Methods in Natural Language Processing* 4360–4370 (Association for Computational Linguistics, 2018).
126. Ryskina, M., Rabinovich, E., Berg-Kirkpatrick, T., Mortensen, D. R. & Tsvetkov, Y. Where new words are born: distributional semantic analysis of neologisms and their semantic neighborhoods. In *Proc. Society for Computation in Linguistics* (Association for Computational Linguistics, 2020).
127. Tomasello, M. First steps toward a usage-based theory of language acquisition. *Cogn. Ling.* **11**, 61–82 (2000).
128. Smith, K., Smith, A. D. & Blythe, R. A. Cross-situational learning: an experimental study of word-learning mechanisms. *Cogn. Sci.* **35**, 480–498 (2011).
129. Watts, D. J. A simple model of global cascades on random networks. *Proc. Natl Acad. Sci. USA* **99**, 5766–5771 (2002).
130. McLeish, K. N. & Oxoby, R. J. Social interactions and the salience of social identity. *J. Econ. Psychol.* **32**, 172–178 (2011).
131. Alhabash, S. & Ma, M. A tale of four platforms: Motivations and uses of facebook, twitter, instagram, and snapchat among college students? *Soc. Media Soc.* **3**, 2056305117691544 (2017).
132. Daganzo, C. F., Gayah, V. V. & Gonzales, E. J. The potential of parsimonious models for understanding large scale transportation systems and answering big picture questions. *EURO J. Transp. Logist.* **1**, 47–65 (2012).
133. Weng, L., Flammini, A., Vespignani, A. & Menczer, F. Competition among memes in a world with limited attention. *Sci. Rep.* **2**, 335 (2012).
134. Shalom, D. E., Sigman, M., Mindlin, G. & Trevisan, M. A. Fading of collective attention shapes the evolution of linguistic variants. *Phys. Rev. E* **100**, 020102 (2019).
135. Bollobás, B. A probabilistic proof of an asymptotic formula for the number of labelled regular graphs. *Eur. J. Comb.* **1**, 311–316 (1980).
136. Lee, S.-I. Developing a bivariate spatial association measure: an integration of pearson's r and moran's i. *J. Geogr. Syst.* **3**, 369–385 (2001).
137. Pimentel, R. S., Niewiadomska-Bugaj, M. & Wang, J.-C. Association of zero-inflated continuous variables. *Stat. Probab. Lett.* **96**, 61–67 (2015).
138. Axelrod, R. *Simulating Social Phenomena* (Springer, 1997).
139. Denevan, W. M. Adaptation, variation, and cultural geography. *Prof. Geogr.* **35**, 399–407 (1983).
140. Wolfram, W. & Schilling-Estes, N. in *The Handbook of Historical Linguistics* Ch. 24 (Blackwell Publishing Oxford, 2003).
141. Rose, G. Cultural geography going viral. *Soc. Cult. Geogr.* **17**, 763–767 (2016).
142. Cisneros, H., Hendricks, D., Clark, J. C. & Fulton, W. *The Texas Triangle: An Emerging Power in the Global Economy* Vol. 27 (Texas A&M University Press, 2021).
143. Gimpel, J. G. & Shaw, D. R. Long distance migration as a two-step sorting process: the resettlement of californians in texas. *Polit. Behav.* 1–28 (2023).
144. Wirth, L. Urbanism as a way of life. *Am.J. Sociol.* **44**, 1–24 (1938).
145. Glenn, N. D. & Hill Jr, L. Rural-urban differences in attitudes and behavior in the united states. *Ann. Am. Acad. Polit. Soc. Sci.* **429**, 36–50 (1977).
146. Meyerhoff, M. & Niedzielski, N. The globalisation of vernacular variation. *J. Socioling.* **7**, 534–555 (2003).
147. Gimpel, J. G., Lovin, N., Moy, B. & Reeves, A. The urban–rural gulf in american political behavior. *Polit. Behav.* **42**, 1343–1368 (2020).
148. Britain, D. Space, Diffusion and Mobility. in *The Handbook of Language Variation and Change* (eds Chambers, J. K., Trudgill, P. & Schilling-Estes, N.) Ch. 22 https://www.wiley.com/en-us/The+Handbook+of+Language+Variation+and+Change%2C+2nd

+Edition-p-9780470659946#tableofcontents-section (Blackwell Publishing, 2004).
149. Cliff, A. & Haggett, P. in *Encyclopedia of Social Measurement* (ed. Kempf-Leonard, K.) (Elsevier, 2005).
150. Labov, W. *Principles of Linguistic Change* (Wiley, 2010).
151. Djelic, M.-L. Social networks and country-to-country transfer: dense and weak ties in the diffusion of knowledge. *Soc. Econ. Rev.* **2**, 341–370 (2004).
152. McAdam, D. & Paulsen, R. Specifying the relationship between social ties and activism. *Am. J. Sociol.* **99**, 640–667 (1993).
153. Centola, D. The spread of behavior in an online social network experiment. *Science* **329**, 1194–1197 (2010).

Acknowledgements

We thank Abigail Z. Jacobs, Nick Ellis, and members of University of Michigan's "Sociolinguistics, Language Contact, and Discourse Analysis" and "Measuring and Modeling Culture" Seminars for their detailed feedback and useful discussion on an early version of this work, leading to changes in experimental design, analysis, and framing. This work has been partially funded by the Air Force Office of Scientific Research under award number FA9550-19-1-0029 and by the National Science Foundation under Grant No 2007251.

Author contributions

A.A., D.J., and D.M.R. designed research; A.A., D.J., and D.M.R. performed research; A.A. and D.J. collected data; A.A. analyzed data; A.A., D.J., and D.M.R. wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s44260-024-00009-9.

Correspondence and requests for materials should be addressed to Aparna Ananthasubramaniam.

Reprints and permissions information is available at http://www.nature.com/reprints