# The Social Dynamics of Language Change in Online Networks

Rahul Goel[1], Sandeep Soni[1], Naman Goyal[1], John Paparrizos[2], Hanna Wallach[3], Fernando Diaz[3], and Jacob Eisenstein[1]

[1] Georgia Institute of Technology, Atlanta, GA, USA
[2] Columbia University, New York, NY, USA
[3] Microsoft Research, New York, NY, USA

**Abstract.** Language change is a complex social phenomenon, revealing pathways of communication and sociocultural influence. But, while language change has long been a topic of study in sociolinguistics, traditional linguistic research methods rely on circumstantial evidence, estimating the direction of change from differences between older and younger speakers. In this paper, we use a data set of several million Twitter users to track language changes in progress. First, we show that language change can be viewed as a form of social influence: we observe complex contagion for phonetic spellings and "netspeak" abbreviations (e.g., *lol*), but not for older dialect markers from spoken language. Next, we test whether specific types of social network connections are more influential than others, using a parametric Hawkes process model. We find that tie strength plays an important role: densely embedded social ties are significantly better conduits of linguistic influence. Geographic locality appears to play a more limited role: we find relatively little evidence to support the hypothesis that individuals are more influenced by geographically local social ties, even in their usage of geographical dialect markers.

## 1 Introduction

Change is a universal property of language. For example, English has changed so much that Renaissance-era texts like *The Canterbury Tales* must now be read in translation. Even contemporary American English continues to change and diversify at a rapid pace—to such an extent that some geographical dialect differences pose serious challenges for comprehensibility [36]. Understanding language change is therefore crucial to understanding language itself, and has implications for the design of more robust natural language processing systems [17].

Language change is a fundamentally social phenomenon [34]. For a new linguistic form to succeed, at least two things must happen: first, speakers (and writers) must come into contact with the new form; second, they must decide to use it. The first condition implies that language change is related to the structure of social networks. If a significant number of speakers are isolated from a potential change, then they are unlikely to adopt it [40]. But mere exposure is not sufficient—we are all exposed to language varieties that are different from our own, yet we nonetheless do not adopt them in our own speech and writing. For example, in the United States, many African American speakers maintain a distinct dialect, despite being immersed in a linguistic environment that differs in many important respects [23,45]. Researchers have made a similar argument for socioeconomic language differences in Britain [49]. In at least some cases, these differences reflect questions of identity: because language is a key constituent in the social construction of group identity, individuals must make strategic choices when deciding whether to adopt new linguistic forms [11,29,33]. By analyzing patterns of language change, we can learn more about the latent structure of social organization: to whom people talk, and how they see themselves.

But, while the basic outline of the interaction between language change and social structure is understood, the fine details are still missing: What types of social network connections are most important for language change? To what extent do considerations of identity affect linguistic differences, particularly in an online context? Traditional sociolinguistic approaches lack the data and the methods for asking such detailed questions about language variation and change.

In this paper, we show that large-scale social media data can shed new light on how language changes propagate through social networks. We use a data set of Twitter users that contains all public messages for several million accounts, augmented with social network and geolocation metadata. This data set makes it possible to track, and potentially explain, every usage of a linguistic variable[4] as it spreads through social media. Overall, we make the following contributions:

- We show that non-standard words are most likely to propagate between individuals who are connected in the Twitter mutual-reply network. This validates the basic approach of using Twitter to measure language change.
- For some classes of non-standard words, we observe complex contagion—i.e., multiple exposures increase the likelihood of adoption. This is particularly true for phonetic spellings and "netspeak" abbreviations. In contrast, non-standard words that originate in speech do not display complex contagion.
- We use a parametric Hawkes process model [26,39] to test whether specific types of social network connections are more influential than others. For some words, we find that densely embedded social ties are significantly better conduits of linguistic influence. This finding suggests that individuals make social evaluations of their exposures to new linguistic forms, and then use these social evaluations to strategically govern their own language use.
- We present an efficient parameter estimation method that uses sparsity patterns in the data to scale to social networks with millions of users.

## 2 Data

Twitter is an online social networking platform. Users post 140-character messages, which appear in their followers' timelines. Because follower ties can be

[4] The basic unit of linguistic differentiation is referred to as a "variable" in the sociolinguistic and dialectological literature [50]. We maintain this terminology here.

asymmetric, Twitter serves multiple purposes: celebrities share messages with millions of followers, while lower-degree users treat Twitter as a more intimate social network for mutual communication [31]. In this paper, we use a large-scale Twitter data set, acquired via an agreement between Microsoft and Twitter. This data set contains all public messages posted between June 2013 and June 2014 by several million users, augmented with social network and geolocation metadata. We excluded retweets, which are explicitly marked with metadata, and focused on messages that were posted in English from within the United States.

### 2.1 Linguistic Markers

The explosive rise in popularity of social media has led to an increase in linguistic diversity and creativity [5,6,9,14,17,27], affecting written language at all levels, from spelling [18] all the way up to grammatical structure [48] and semantic meaning across the lexicon [25,30]. Here, we focus on the most easily observable and measurable level: variation and change in the use of individual words.

We take as our starting point words that are especially characteristic of eight cities in the United States. We chose these cities to represent a wide range of geographical regions, population densities, and demographics. We identified the following words as geographically distinctive markers of their associated cities, using SAGE [20]. Specifically, we followed the approach previously used by Eisenstein to identify community-specific terms in textual corpora [19].[5]

**Atlanta:** *ain* (phonetic spelling of *ain't*), *dese* (phonetic spelling of *these*), *yeen* (phonetic spelling of *you ain't*);
**Baltimore:** *ard* (phonetic spelling of *alright*), *inna* (phonetic spelling of *in a* and *in the*), *lls* (*laughing like shit*), *phony* (fake);
**Charlotte:** *cookout*;
**Chicago:** *asl* (phonetic spelling of *as hell*, typically used as an intensifier on Twitter[6]), *mfs* (*motherfuckers*);
**Los Angeles:** *graffiti*, *tfti* (*thanks for the information*);
**Philadelphia:** *ard* (phonetic spelling of *alright*), *ctfuu* (expressive lengthening of *ctfu*, an abbreviation of *cracking the fuck up*), *jawn* (generic noun);
**San Francisco:** *hella* (an intensifier);
**Washington D.C.:** *inna* (phonetic spelling of *in a* and *in the*), *lls* (*laughing like shit*), *stamp* (an exclamation indicating emphasis).[7]

Linguistically, we can divide these words into three main classes:

[5] After running SAGE to identify words with coefficients above 2.0, we manually removed hashtags, named entities, non-English words, and descriptions of events.

[6] Other sources, such as `http://urbandictionary.com`, report *asl* to be an abbreviation of *age, sex, location?* However, this definition is not compatible with typical usage on Twitter, e.g., *currently hungry asl* or *that movie was funny asl.*

[7] *ard*, *inna*, and *lls* appear on multiple cities' lists. These words are characteristic of the neighboring cities of Baltimore, Philadelphia, and Washington D.C.

**Lexical words:** The origins of *cookout*, *graffiti*, *hella*, *phony*, and *stamp* can almost certainly be traced back to spoken language. Some of these words (e.g., *cookout* and *graffiti*) are known to all fluent English speakers, but are preferred in certain cities simply as a matter of topic. Other words (e.g., *hella* [12] and *jawn* [3]) are dialect markers that are not widely used outside their regions of origin, even after several decades of use in spoken language.

**Phonetic spellings:** *ain*, *ard*, *asl*, *inna*, and *yeen* are non-standard spellings that are based on phonetic variation by region, demographics, or situation.

**Abbreviations:** *ctfuu*, *lls*, *mfs*, and *tfti* are phrasal abbreviations. These words are interesting because they are fundamentally textual. They are unlikely to have come from spoken language, and are intrinsic to written social media.

Several of these words were undergoing widespread growth in popularity around the time period spanned by our data set. For example, the frequencies of *ard*, *asl*, *hella*, and *tfti* more than tripled between 2012 and 2013. Our main research question is whether and how these words spread through Twitter. For example, lexical words are mainly transmitted through speech. We would expect their spread to be only weakly correlated with the Twitter social network. In contrast, abbreviations are fundamentally textual in nature, so we would expect their spread to correlate much more closely with the Twitter social network.

### 2.2 Social network

To focus on communication between peers, we constructed a social network of mutual replies between Twitter users. Specifically, we created a graph in which there is a node for each user in the data set. We then placed an undirected edge between a pair of users if each replied to the other by beginning a message with their username. Our decision to use the reply network (rather than the follower network) was a pragmatic choice: the follower network is not widely available. However, the reply network is also well supported by previous research. For example, Huberman *et al.* argue that Twitter's mention network is more socially meaningful than its follower network: although users may follow thousands of accounts, they interact with a much more limited set of users [28], bounded by a constant known as Dunbar's number [15]. Finally, we restricted our focus to mutual replies because there are a large number of unrequited replies directed at celebrities. These replies do not indicate a meaningful social connection.

We compared our mutual-reply network with two one-directional "in" and "out" networks, in which all public replies are represented by directed edges. The degree distributions of these networks are depicted in Figure 1. As expected, there are a few celebrities with very high in-degrees, and a maximum in-degree of $20,345$. In contrast, the maximum degree in our mutual-reply network is 248.

### 2.3 Geography

In order to test whether geographically local social ties are a significant conduit of linguistic influence, we obtained geolocation metadata from Twitter's location

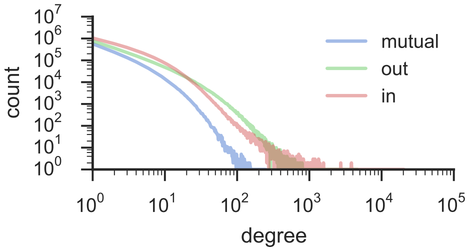

**Fig. 1.** Degree distributions for our mutual-reply network and "in" and "out" networks.

field. This field is populated via a combination of self reports and GPS tagging. We aggregated metadata across each user's messages, so that each user was geolocated to the city from which they most commonly post messages. Overall, our data set contains 4.35 million geolocated users, of which 589,562 were geolocated to one of the eight cities listed in § 2.1. We also included the remaining users in our data set, but were not able to account for their geographical location.

Researchers have previously shown that social network connections in online social media tend to be geographically assortative [7,46]. Our data set is consistent with this finding: for 94.8% of mutual-reply dyads in which both users were geolocated to one of the eight cities listed in § 2.1, they were both geolocated to the same city. This assortativity motivates our decision to estimate separate influence parameters for local and non-local social connections (see § 5.1).

## 3 Language Change as Social Influence

Our main research goal is to test whether and how geographically distinctive linguistic markers spread through Twitter. With this goal in mind, our first question is whether the adoption of these markers can be viewed as a form of COMPLEX CONTAGION. To answer this question, we computed the fraction of users who used one of the words listed in § 2.1 after being exposed to that word by one of their social network connections. Formally, we say that user $i$ EXPOSED user $j$ to word $w$ at time $t$ if and only if the following conditions hold: $i$ used $w$ at time $t$; $j$ had not used $w$ before time $t$; the social network connection $i \leftrightarrow j$ was formed before time $t$. We define the INFECTION RISK for word $w$ to be the number of users who use word $w$ after being exposed divided by the total number of users who were exposed. To consider the possibility that multiple exposures have a greater impact on the infection risk, we computed the infection risk after exposures across one, two, and three or more distinct social network connections.

The words' infection risks cannot be interpreted directly because relational autocorrelation can also be explained by homophily and external confounds. For example, geographically distinctive non-standard language is more likely to be used by young people [44], and online social network connections are assortative by age [2]. Thus, a high infection risk can also be explained by the confound of age. We therefore used the shuffle test proposed by Anagnostopoulos *et al.* [4],

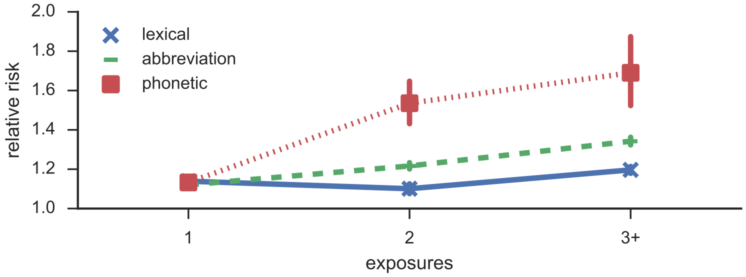

**Fig. 2.** Relative infection risks for words in each of the three linguistic classes defined in § 2.1. The figure depicts 95% confidence intervals, computed using the shuffle test [4].

which compares the observed infection risks to infection risks under the null hypothesis that event timestamps are independent. The null hypothesis infection risks are computed by randomly permuting the order of word usage events. If the observed infection risks are substantially higher than the infection risks computed using the permuted data, then this is compatible with social influence.[8]

Figure 2 depicts the ratios between the words' observed infection risks and the words' infection risks under the null hypothesis, after exposures across one, two, and three or more distinct connections. We computed 95% confidence intervals across the words and across the permutations used in the shuffle test. For all three linguistic classes defined in § 2.1, the risk ratio for even a single exposure is significantly greater than one, suggesting the existence of social influence. The risk ratio for a single exposure is nearly identical across the three classes. For phonetic spellings and abbreviations, the risk ratio grows with the number of exposures. This pattern suggests that words in these classes exhibit COMPLEX CONTAGION—i.e., multiple exposures increase the likelihood of adoption [13]. In contrast, the risk ratio for lexical words remains the same as the number of exposures increases, suggesting that these words spread by simple contagion.

Complex contagion has been linked to a range of behaviors, from participation in collective political action to adoption of avant garde fashion [13]. A common theme among these behaviors is that they are not cost-free, particularly if the behavior is not legitimated by widespread adoption. In the case of linguistic markers intrinsic to social media, such as phonetic spellings and abbreviations, adopters risk negative social evaluations of their linguistic competency, as well as their cultural authenticity [47]. In contrast, lexical words are already well known from spoken language and are thus less socially risky. This difference may explain why we do not observe complex contagion for lexical words.

[8] The shuffle test assumes that the likelihood of two users forming a social network connection does not change over time. Researchers have proposed a test [32] that removes this assumption; we will scale this test to our data set in future work.

## 4 Social Evaluation of Language Variation

In the previous section, we showed that geographically distinctive linguistic markers spread through Twitter, with evidence of complex contagion for phonetic spellings and abbreviations. But, does each social network connection contribute equally? Our second question is therefore whether (1) strong ties and (2) geographically local ties exert greater linguistic influence than other ties. If so, users must socially evaluate the information they receive from these connections, and judge it to be meaningful to their linguistic self-presentation. In this section, we outline two hypotheses regarding their relationships to linguistic influence.

### 4.1 Tie Strength

Social networks are often characterized in terms of strong and weak ties [22,40], with strong ties representing more important social relationships. Strong ties are often densely embedded, meaning that the nodes in question share many mutual friends; in contrast, weak ties often bridge disconnected communities. Bakshy *et al.* investigated the role of weak ties in information diffusion, through resharing of URLs on Facebook [8]. They found that URLs shared across strong ties are more likely to be reshared. However, they also found that weak ties play an important role, because users tend to have more weak ties than strong ties, and because weak ties are more likely to be a source of new information. In some respects, language change is similar to traditional information diffusion scenarios, such as resharing of URLs. But, in contrast, language connects with personal identity on a much deeper level than a typical URL. As a result, strong, deeply embedded ties may play a greater role in enforcing community norms.

We quantify tie strength in terms of EMBEDDEDNESS. Specifically, we use the normalized mutual friends metric introduced by Adamic and Adar [1]:

$$s_{i,j} = \sum_{k \in \mathit{\Gamma}(i) \cap \mathit{\Gamma}(j)} \frac{1}{\log\left(\#|\mathit{\Gamma}(k)|\right)}, \tag{1}$$

where, in our setting, $\mathit{\Gamma}(i)$ is the set of users connected to $i$ in the Twitter mutual-reply network and $\#|\mathit{\Gamma}(i)|$ is the size of this set. This metric rewards dyads for having many mutual friends, but counts mutual friends more if their degrees are low—a high-degree mutual friend is less informative than one with a lower-degree. Given this definition, we can form the following hypothesis:

**H1** The linguistic influence exerted across ties with a high embeddedness value $s_{i,j}$ will be greater than the linguistic influence exerted across other ties.

### 4.2 Geographic Locality

An open question in sociolinguistics is whether and how local COVERT PRESTIGE—i.e., the positive social evaluation of non-standard dialects—affects the adoption of new linguistic forms [49]. Speakers often explain their linguistic choices in terms of their relationship with their local identity [16], but this may be a post-hoc rationalization made by people whose language is affected by factors beyond their control. Indeed, some sociolinguists have cast doubt on the role of "local games" in affecting the direction of language change [35].

The theory of covert prestige suggests that geographically local social ties are more influential than non-local ties. We do not know of any prior attempts to test this hypothesis quantitatively. Although researchers have shown that local linguistic forms are more likely to be used in messages that address geographically local friends [43], they have not attempted to measure the impact of exposure to these forms. This lack of prior work may be because it is difficult to obtain relevant data, and to make reliable inferences from such data. For example, there are several possible explanations for the observation that people often use similar language to that of their geographical neighbors. One is exposure: even online social ties tend to be geographically assortative [2], so most people are likely to be exposed to local linguistic forms through local ties. Alternatively, the causal relation may run in the reverse direction, with individuals preferring to form social ties with people whose language matches their own. In the next section, we describe a model that enables us to tease apart the roles of geographic assortativity and local influence, allowing us to test the following hypothesis:

**H2** The influence toward geographically distinctive linguistic markers is greater when exerted across geographically local ties than across other ties.

We note that this hypothesis is restricted in scope to geographically distinctive words. We do not consider the more general hypothesis that geographically local ties are more influential for all types of language change, such as change involving linguistic variables that are associated with gender or socioeconomic status.

## 5 Language Change as a Self-exciting Point Process

To test our hypotheses about social evaluation, we require a more sophisticated modeling tool than the simple counting method described in § 3. In this section, rather than asking whether a user was previously exposed to a word, we ask by whom, in order to compare the impact of exposures across different types of social network connections. We also consider temporal properties. For example, if a user adopts a new word, should we credit this to an exposure from a weak tie in the past hour, or to an exposure from a strong tie in the past day?

Following a probabilistic modeling approach, we treated our Twitter data set as a set of cascades of timestamped events, with one cascade for each of the geographically distinctive words described in § 2.1. Each event in a word's cascade corresponds to a tweet containing that word. We modeled each cascade as a probabilistic process, and estimated the parameters of this process. By comparing nested models that make progressively finer distinctions between social network connections, we were able to quantitatively test our hypotheses.

Our modeling framework is based on a Hawkes process [26]—a specialization of an inhomogeneous Poisson process—which explains a cascade of timestamped events in terms of influence parameters. In a temporal setting, an inhomogeneous Poisson process says that the number of events $y_{t_1,t_2}$ between $t_1$ and $t_2$ is drawn from a Poisson distribution, whose parameter is the area under a time-varying INTENSITY FUNCTION over the interval defined by $t_1$ and $t_2$:

$$y_{t_1,t_2} \sim \text{Poisson}\left(\mathit{\Lambda}(t_1, t_2)\right)) \tag{2}$$

where

$$\mathit{\Lambda}(t_1, t_2) = \int_{t_1}^{t_2} \lambda(t)\ \mathrm{d}t. \tag{3}$$

Since the parameter of a Poisson distribution must be non-negative, the intensity function must be constrained to be non-negative for all possible values of $t$.

A Hawkes process is a self-exciting inhomogeneous Poisson process, where the intensity function depends on previous events. If we have a cascade of $N$ events $\{t_n\}_{n=1}^{N}$, where $t_n$ is the timestamp of event $n$, then the intensity function is

$$\lambda(t) = \mu_t + \sum_{t_n < t} \alpha\, \kappa(t - t_n), \tag{4}$$

where $\mu_t$ is the base intensity at time $t$, $\alpha$ is an influence parameter that captures the influence of previous events, and $\kappa(\cdot)$ is a time-decay kernel.

We can extend this framework to vector observations $\boldsymbol{y}_{t_1,t_2} = (y_{t_1,t_2}^{(1)}, \ldots, y_{t_1,t_2}^{(M)})$ and intensity functions $\boldsymbol{\lambda}(t) = (\lambda^{(1)}(t), \ldots, \lambda^{(M)}(t))$, where, in our setting, $M$ is the total number of users in our data set. If we have a cascade of $N$ events $\{(t_n, m_n)\}_{n=1}^{N}$, where $t_n$ is the timestamp of event $n$ and $m_n \in \{1, \ldots, M\}$ is the source of event $n$, then the intensity function for user $m' \in \{1, \ldots, M\}$ is

$$\lambda^{(m')}(t) = \mu_t^{(m')} + \sum_{t_n < t} \alpha_{m_n \to m'} \kappa(t - t_n), \tag{5}$$

where $\mu_t^{(m')}$ is the base intensity for user $m'$ at time $t$, $\alpha_{m_n \to m'}$ is a pairwise influence parameter that captures the influence of user $m_n$ on user $m'$, and $\kappa(\cdot)$ is a time-decay kernel. Throughout our experiments, we used an exponential decay kernel $\kappa(\mathit{\Delta} t) = e^{-\gamma \mathit{\Delta} t}$. We set the hyperparameter $\gamma$ so that $\kappa(1 \text{ hour}) = e^{-1}$.

Researchers usually estimate all $M^2$ influence parameters of a Hawkes process (e.g., [38,51]). However, in our setting, $M > 10^6$, so there are $O(10^{12})$ influence parameters. Estimating this many parameters is computationally and statistically intractable, given that our data set includes only $O(10^5)$ events (see the $x$-axis of Figure 3 for event counts for each word). Moreover, directly estimating these parameters does not enable us to quantitatively test our hypotheses.

### 5.1 Parametric Hawkes Process

Instead of directly estimating all $O(M^2)$ pairwise influence parameters, we used Li and Zha's parametric Hawkes process [39]. This model defines each pairwise influence parameter in terms of a linear combination of pairwise features:

$$\alpha_{m \to m'} = \boldsymbol{\theta}^\top \boldsymbol{f}(m \to m'), \tag{6}$$

where $\boldsymbol{f}(m \to m')$ is a vector of features that describe the relationship between users $m$ and $m'$. Thus, we only need to estimate the feature weights $\boldsymbol{\theta}$ and the base intensities. To ensure that the intensity functions $\lambda^{(1)}(t), \ldots, \lambda^{(M)}(t)$ are non-negative, we must assume that $\boldsymbol{\theta}$ and the base intensities are non-negative.

We chose a set of four binary features that would enable us to test our hypotheses about the roles of different types of social network connections:

**F1 Self-activation:** This feature fires when $m' = m$. We included this feature to capture the scenario where using a word once makes a user more likely to use it again, perhaps because they are adopting a non-standard style.

**F2 Mutual reply:** This feature fires if the dyad $(m, m')$ is in the Twitter mutual-reply network described in § 2.2. We also used this feature to define the remaining two features. By doing this, we ensured that features F2, F3, and F4 were (at least) as sparse as the mutual-reply network.

**F3 Tie strength:** This feature fires if the dyad $(m, m')$ is in in the Twitter mutual-reply network, and the Adamic-Adar value for this dyad is especially high. Specifically, we require that the Adamic-Adar value be in the 90th percentile among all dyads where at least one user has used the word in question. Thus, this feature picks out the most densely embedded ties.

**F4 Local:** This feature fires if the dyad $(m, m')$ is in the Twitter mutual-reply network, and the users were geolocated to the same city, and that city is one of the eight cities listed in § 2. For other dyads, this feature returns zero. Thus, this feature picks out a subset of the geographically local ties.

In § 6, we describe how we used these features to construct a set of nested models that enabled us to test our hypotheses. In the remainder of this section, we provide the mathematical details of our parameter estimation method.

### 5.2 Objective Function

We estimated the parameters using constrained maximum likelihood. Given a cascade of events $\{(t_n, m_n)\}_{n=1}^N$, the log likelihood under our model is

$$\mathcal{L} = \sum_{n=1}^{N} \log \lambda^{(m_n)}(t_n) - \sum_{m=1}^{M} \int_0^T \lambda^{(m)}(t)\ \mathrm{d}t, \tag{7}$$

where $T$ is the temporal endpoint of the cascade. Substituting in the complete definition of the per-user intensity functions from Equation 5 and Equation 6,

$$\mathcal{L} = \sum_{n=1}^{N} \log \left( \mu_{t_n}^{(m_n)} + \sum_{t_{n'} < t_n} \boldsymbol{\theta}^\top \boldsymbol{f}(m_{n'} \to m_n)\, \kappa(t_n - t_{n'}) \right) - \sum_{m'=1}^{M} \int_0^T \left( \mu_t^{(m')} + \sum_{t_{n'} < t} \boldsymbol{\theta}^\top \boldsymbol{f}(m_{n'} \to m')\, \kappa(t - t_{n'}) \right) \mathrm{d}t. \tag{8}$$

If the base intensities are constant with respect to time, then

$$\mathcal{L} = \sum_{n=1}^{N} \log \left( \mu^{(m_n)} + \sum_{t_{n'} < t_n} \boldsymbol{\theta}^{\top} \boldsymbol{f}(m_{n'} \rightarrow m_n)\, \kappa(t_n - t_{n'}) \right) - \sum_{m'=1}^{M} \left( T\mu^{(m')} + \sum_{n=1}^{N} \boldsymbol{\theta}^{\top} \boldsymbol{f}(m_n \rightarrow m')\, (1 - \kappa(T - t_n)) \right), \tag{9}$$

where the second term includes a sum over all events $n = \{1, \ldots, N\}$ that contibute to the final intensity $\lambda^{(m')}(T)$. To ease computation, however, we can rearrange the second term around the source $m$ rather than the recipient $m'$:

$$\mathcal{L} = \sum_{n=1}^{N} \log \left( \mu^{(m_n)} + \sum_{t_{n'} < t_n} \boldsymbol{\theta}^{\top} \boldsymbol{f}(m_{n'} \rightarrow m_n)\, \kappa(t_n - t_{n'}) \right) - \sum_{m=1}^{M} \left( T\mu^{(m)} + \sum_{\{n : m_n = m\}} \boldsymbol{\theta}^{\top} \boldsymbol{f}(m \rightarrow \star)\, (1 - \kappa(T - t_n)) \right), \tag{10}$$

where we have introduced an aggregate feature vector $\boldsymbol{f}(m \rightarrow \star) = \sum_{m'=1}^{M} \boldsymbol{f}(m \rightarrow m')$. Because the sum $\sum_{\{n : m_n = m'\}} \boldsymbol{f}(m' \rightarrow \star)\, \kappa(T - t_n)$ does not involve either $\boldsymbol{\theta}$ or $\mu^{(1)}, \ldots, \mu^{(M)}$, we can pre-compute it. Moreover, we need to do so only for users $m \in \{1, \ldots, M\}$ for whom there is at least one event in the cascade.

A Hawkes process defined in terms of Equation 5 has a log likelihood that is convex in the pairwise influence parameters and the base intensities. For a parametric Hawkes process, $\alpha_{m \rightarrow m'}$ is an affine function of $\boldsymbol{\theta}$, so, by composition, the log likelihood is convex in $\boldsymbol{\theta}$ and remains convex in the base intensities.

### 5.3 Gradients

The first term in the log likelihood and its gradient contains a nested sum over events, which appears to be quadratic in the number of events. However, we can use the exponential decay of the kernel $\kappa(\cdot)$ to approximate this term by setting a threshold $\tau^{\star}$ such that $\kappa(t_n - t_{n'}) = 0$ if $t_n - t_{n'} \geq \tau^{\star}$. For example, if we set $\tau^{\star} = 24$ hours, then we approximate $\kappa(\tau^{\star}) = 3 \times 10^{-11} \approx 0$. This approximation makes the cost of computing the first term linear in the number of events.

The second term is linear in the number of social network connections and linear in the number of events. Again, we can use the exponential decay of the kernel $\kappa(\cdot)$ to approximate $\kappa(T - t_n) \approx 0$ for $T - t_n \geq \tau^{\star}$, where $\tau^{\star} = 24$ hours. This approximation means that we only need to consider a small number of tweets near temporal endpoint of the cascade. For each user, we also pre-computed $\sum_{\{n : m_n = m'\}} \boldsymbol{f}(m' \rightarrow \star)\, \kappa(T - t_n)$. Finally, both terms in the log likelihood and its gradient can also be trivially parallelized over users $m = \{1, \ldots, M\}$.

For a Hawkes process defined in terms of Equation 5, Ogata showed that additional speedups can be obtained by recursively pre-computing a set of aggregate messages for each dyad $(m, m')$. Each message represents the events from user $m$ that may influence user $m'$ at the time $t_i^{(m')}$ of their $i^{\text{th}}$ event [42]:

$$\begin{aligned}
&R^{(i)}_{m\to m'} \\
&\quad = \begin{cases} \kappa(t_i^{(m')} - t_{i-1}^{(m')})\, R^{(i-1)}_{m\to m'} + \sum_{t_{i-1}^{(m')} \le t_j^{(m)} \le t_i^{(m')}} \kappa(t_i^{(m')} - t_j^{(m)}) & m \neq m' \\ \kappa(t_i^{(m')} - t_{i-1}^{(m')}) \times (1 + R^{(i-1)}_{m\to m'}) & m = m'. \end{cases}
\end{aligned}$$

These aggregate messages do not involve the feature weights $\boldsymbol{\theta}$ or the base intensities, so they can be pre-computed and reused throughout parameter estimation.

For a parametric Hawkes process, it is not necessary to compute a set of aggregate messages for each dyad. It is sufficient to compute a set of aggregate messages for each possible configuration of the features. In our setting, there are only four binary features, and some combinations of features are impossible.

Because the words described in § 2.1 are relatively rare, most of the users in our data set never used them. However, it is important to include these users in the model. Because they did not adopt these words, despite being exposed to them by users who did, their presence exerts a negative gradient on the feature weights. Moreover, such users impose a minimal cost on parameter estimation because they need to be considered only when pre-computing feature counts.

### 5.4 Coordinate Ascent

We optimized the log likelihood with respect to the feature weights $\boldsymbol{\theta}$ and the base intensities. Because the log likelihood decomposes over users, each base intensity $\mu^{(m)}$ is coupled with only the feature weights and not with the other base intensities. Jointly estimating all parameters is inefficient because it does not exploit this structure. We therefore used a coordinate ascent procedure, alternating between updating $\boldsymbol{\theta}$ and the base intensities. As explained in § 5.1, both $\boldsymbol{\theta}$ and the base intensities must be non-negative to ensure that intensity functions are also non-negative. At each stage of the coordinate ascent, we performed constrained optimization using the active set method of MATLAB's `fmincon` function.

## 6 Results

We used a separate set of parametric Hawkes process models for each of the geographically distinctive linguistic markers described in § 2.1. Specifically, for each word, we constructed a set of nested models by first creating a baseline model using features F1 (self-activation) and F2 (mutual reply) and then adding in each of the experimental features—i.e., F3 (tie strength) and F4 (local).

We tested hypothesis H1 (strong ties are more influential) by comparing the goodness of fit for feature set F1+F2+F3 to that of feature set F1+F2. Similarly, we tested H2 (geographically local ties are more influential) by comparing the goodness of fit for feature set F1+F2+F4 to that of feature set F1+F2.

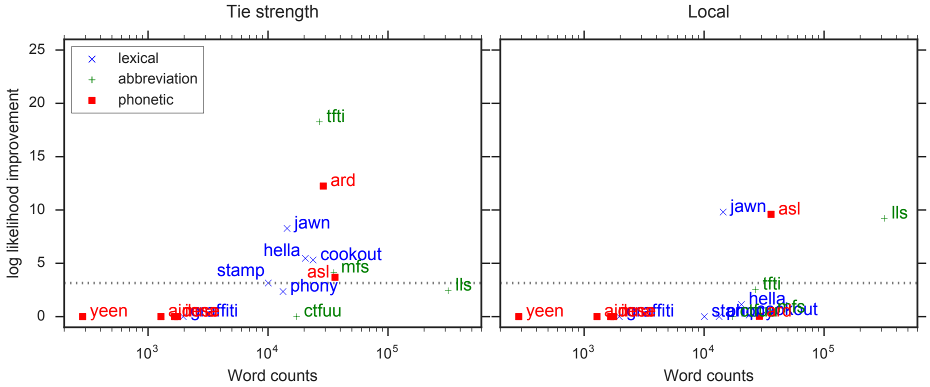

**Fig. 3.** Improvement in goodness of fit from adding in features F3 (tie strength) and F4 (local). The dotted line corresponds to the threshold for statistical significance at $p < 0.05$ using a likelihood ratio test with the Benjamini-Hochberg correction.

In Figure 3, we show the improvement in goodness of fit from adding in features F3 and F4.[9] Under the null hypothesis, the log of the likelihood ratio follows a $\chi^2$ distribution with one degree of freedom, because the models differ by one parameter. Because we performed thirty-two hypothesis tests (sixteen words, two features), we needed to adjust the significance thresholds to correct for multiple comparisons. We did this using the Benjamini-Hochberg procedure [10].

Features F3 and F4 did not improve the goodness of fit for less frequent words, such as *ain*, *graffiti*, and *yeen*, which occur fewer than $10^4$ times. Below this count threshold, there is not enough data to statistically distinguish between different types of social network connections. However, above this count threshold, adding in F3 (tie strength) yielded a statistically significant increase in goodness of fit for *ard*, *asl*, *cookout*, *hella*, *jawn*, *mfs*, and *tfti*. This finding provides evidence in favor of hypothesis H1—that the linguistic influence exerted across densely embedded ties is greater than the linguistic influence exerted across other ties.

In contrast, adding in F4 (local) only improved goodness of fit for three words: *asl*, *jawn*, and *lls*. We therefore conclude that support for hypothesis H2—that the linguistic influence exerted across geographically local ties is greater than the linguistic influence across than across other ties—is limited at best.

In § 3 we found that phonetic spellings and abbreviations exhibit complex contagion, while lexical words do not. Here, however, we found no such systematic differences between the three linguistic classes. Although we hypothesize that lexical words propagate mainly outside of social media, we nonetheless see that when these words do propagate across Twitter, their adoption is modulated by tie strength, as is the case for phonetic spellings and abbreviations.

[9] We also compared the full feature set—i.e., F1+F2+F3+F4—to feature set F1+F2+F3 and feature set F1+F2+F4. The results were almost identical, indicating that F3 (tie strength) and F4 (local) provide complementary information.

## 7 Discussion

Our results in § 3 demonstrate that language change in social media can be viewed as a form of information diffusion across a social network. Moreover, this diffusion is modulated by a number of sociolinguistic factors. For non-lexical words, such as phonetic spellings and abbreviations, we find evidence of complex contagion: the likelihood of their adoption increases with the number of exposures. For both lexical and non-lexical words, we find evidence that the linguistic influence exerted across densely embedded ties is greater than the linguistic influence exerted across other ties. In contrast, we find no evidence to support the hypothesis that geographically local ties are more influential.

Overall, these findings indicate that language change is not merely a process of random diffusion over an undifferentiated social network, as proposed in many simulation studies [21,24,41]. Rather, some social network connections matter more than others, and social judgments have a role to play in modulating language change. In turn, this conclusion provides large-scale quantitative support for earlier findings from ethnographic studies. A logical next step would be to use these insights to design more accurate simulation models, which could be used to reveal long-term implications for language variation and change.

Extending our study beyond North America is a task for future work. Social networks vary dramatically across cultures, with traditional societies tending toward networks with fewer but stronger ties [40]. The social properties of language variation in these societies may differ as well. Another important direction for future work is to determine the impact of exogenous events, such as the appearance of new linguistic forms in mass media. Exogeneous events pose potential problems for estimating both infection risks and social influence. However, it may be possible to account for these events by incorporating additional data sources, such as search trends. Finally, we plan to use our framework to study the spread of terminology and ideas through networks of scientific research articles. Here too, authors may make socially motivated decisions to adopt specific terms and ideas [37]. The principles behind these decisions might therefore be revealed by an analysis of linguistic events propagating over a social network.

*Acknowledgments* Thanks to the reviewers for their feedback, to Márton Karsai for suggesting the infection risk analysis, and to Le Song for discussing Hawkes processes. John Paparrizos is an Alexander S. Onassis Foundation Scholar. This research was supported by the National Science Foundation under awards IIS-1111142 and RI-1452443, by the National Institutes of Health under award number R01-GM112697-01, and by the Air Force Office of Scientific Research.

# References

1. Adamic, L.A., Adar, E.: Friends and neighbors on the web. Social networks 25(3), 211–230 (2003)
2. Al Zamal, F., Liu, W., Ruths, D.: Homophily and latent attribute inference: Inferring latent attributes of Twitter users from neighbors. In: Proceedings of the International Conference on Web and Social Media (ICWSM). pp. 387–390 (2012)
3. Alim, H.S.: Hip hop nation language. In: Duranti, A. (ed.) Linguistic Anthropology: A Reader, pp. 272–289. Wiley-Blackwell, Malden, MA (2009)
4. Anagnostopoulos, A., Kumar, R., Mahdian, M.: Influence and correlation in social networks. In: Proceedings of Knowledge Discovery and Data Mining (KDD). pp. 7–15 (2008)
5. Androutsopoulos, J.: Language change and digital media: a review of conceptions and evidence. In: Coupland, N., Kristiansen, T. (eds.) Standard Languages and Language Standards in a Changing Europe. Novus, Oslo (2011)
6. Anis, J.: Neography: Unconventional spelling in French SMS text messages. In: Danet, B., Herring, S.C. (eds.) The Multilingual Internet: Language, Culture, and Communication Online, pp. 87–115. Oxford University Press (2007)
7. Backstrom, L., Sun, E., Marlow, C.: Find me if you can: improving geographical prediction with social and spatial proximity. In: Proceedings of the Conference on World-Wide Web (WWW). pp. 61–70 (2010)
8. Bakshy, E., Rosenn, I., Marlow, C., Adamic, L.: The role of social networks in information diffusion. In: Proceedings of the Conference on World-Wide Web (WWW). pp. 519–528. Lyon, France (2012)
9. Baldwin, T., Cook, P., Lui, M., MacKinlay, A., Wang, L.: How noisy social media text, how diffrnt social media sources. In: Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP 2013). pp. 356–364 (2013)
10. Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological) pp. 289–300 (1995)
11. Bucholtz, M., Hall, K.: Identity and interaction: A sociocultural linguistic approach. Discourse studies 7(4-5), 585–614 (2005)
12. Bucholtz, M., Bermudez, N., Fung, V., Edwards, L., Vargas, R.: Hella Nor Cal or totally So Cal? The perceptual dialectology of California. Journal of English Linguistics 35(4), 325–352 (2007)
13. Centola, D., Macy, M.: Complex contagions and the weakness of long ties. American journal of Sociology 113(3), 702–734 (2007)
14. Crystal, D.: Language and the Internet. Cambridge University Press, second edn. (sep 2006)
15. Dunbar, R.I.: Neocortex size as a constraint on group size in primates. Journal of Human Evolution 22(6), 469–493 (1992)
16. Eckert, P.: Linguistic variation as social practice. Blackwell (2000)
17. Eisenstein, J.: What to do about bad language on the internet. In: Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL). pp. 359–369 (2013)
18. Eisenstein, J.: Systematic patterning in phonologically-motivated orthographic variation. Journal of Sociolinguistics 19, 161–188 (2015)
19. Eisenstein, J.: Written dialect variation in online social media. In: Boberg, C., Nerbonne, J., Watt, D. (eds.) Handbook of Dialectology. Wiley (2016)

20. Eisenstein, J., Ahmed, A., Xing, E.P.: Sparse additive generative models of text. In: Proceedings of the International Conference on Machine Learning (ICML). pp. 1041–1048 (2011)
21. Fagyal, Z., Swarup, S., Escobar, A.M., Gasser, L., Lakkaraju, K.: Centers and peripheries: Network roles in language change. Lingua 120(8), 2061–2079 (2010)
22. Granovetter, M.S.: The strength of weak ties. American journal of sociology pp. 1360–1380 (1973)
23. Green, L.J.: African American English: A Linguistic Introduction. Cambridge University Press, Cambridge, U.K. (2002)
24. Griffiths, T.L., Kalish, M.L.: Language evolution by iterated learning with bayesian agents. Cognitive Science 31(3), 441–480 (2007)
25. Hamilton, W.L., Leskovec, J., Jurafsky, D.: Diachronic word embeddings reveal statistical laws of semantic change. In: Proceedings of the Association for Computational Linguistics (ACL). Berlin (2016)
26. Hawkes, A.G.: Spectra of some self-exciting and mutually exciting point processes. Biometrika 58(1), 83–90 (1971)
27. Herring, S.C.: Grammar and electronic communication. In: Chapelle, C.A. (ed.) The Encyclopedia of Applied Linguistics. Wiley (2012)
28. Huberman, B., Romero, D.M., Wu, F.: Social networks that matter: Twitter under the microscope. First Monday 14(1) (2008)
29. Johnstone, B., Bhasin, N., Wittkofski, D.: "Dahntahn" Pittsburgh: Monophthongal /aw/ and Representations of Localness in Southwestern Pennsylvania. American Speech 77(2), 148–176 (2002)
30. Kulkarni, V., Al-Rfou, R., Perozzi, B., Skiena, S.: Statistically significant detection of linguistic change. In: Proceedings of the Conference on World-Wide Web (WWW). pp. 625–635 (2015)
31. Kwak, H., Lee, C., Park, H., Moon, S.: What is Twitter, a social network or a news media? In: Proceedings of the Conference on World-Wide Web (WWW). pp. 591–600 (2010)
32. La Fond, T., Neville, J.: Randomization tests for distinguishing social influence and homophily effects. In: Proceedings of the Conference on World-Wide Web (WWW). pp. 601–610 (2010)
33. Labov, W.: The social motivation of a sound change. Word 19(3), 273–309 (1963)
34. Labov, W.: Principles of Linguistic Change, vol. 2: Social Factors. Wiley-Blackwell (2001)
35. Labov, W.: Review of Linguistic Variation as Social Practice, by Penelope Eckert. Language in Society 31, 277–284 (4 2002)
36. Labov, W.: Principles of Linguistic Change, vol. 3: Cognitive and Cultural Factors. Wiley-Blackwell (2011)
37. Latour, B., Woolgar, S.: Laboratory life: The construction of scientific facts. Princeton University Press (2013)
38. Li, L., Deng, H., Dong, A., Chang, Y., Zha, H.: Identifying and labeling search tasks via query-based Hawkes processes. In: Proceedings of Knowledge Discovery and Data Mining (KDD). pp. 731–740 (2014)
39. Li, L., Zha, H.: Learning parametric models for social infectivity in multidimensional hawkes processes. In: Proceedings of the National Conference on Artificial Intelligence (AAAI) (2015)
40. Milroy, L., Milroy, J.: Social network and social class: Toward an integrated sociolinguistic model. Language in society 21(01), 1–26 (1992)
41. Niyogi, P., Berwick, R.C.: A dynamical systems model for language change. Complex Systems 11(3), 161–204 (1997)

42. Ogata, Y.: On lewis' simulation method for point processes. Information Theory, IEEE Transactions on 27(1), 23–31 (1981)
43. Pavalanathan, U., Eisenstein, J.: Audience-modulated variation in online social media. American Speech 90(2) (May 2015)
44. Pavalanathan, U., Eisenstein, J.: Confounds and consequences in geotagged twitter data. In: Proceedings of Empirical Methods for Natural Language Processing (EMNLP) (September 2015)
45. Rickford, J.R.: Geographical diversity, residential segregation, and the vitality of african american vernacular english and its speakers. Transforming Anthropology 18(1), 28–34 (2010)
46. Sadilek, A., Kautz, H., Bigham, J.P.: Finding your friends and following them to where you are. In: Proceedings of the Conference on Web Search and Data Mining (WSDM). pp. 723–732 (2012)
47. Squires, L.: Enregistering internet language. Language in Society 39, 457–492 (2010)
48. Tagliamonte, S.A., Denis, D.: Linguistic ruin? LOL! Instant messaging and teen language. American Speech 83(1), 3–34 (2008)
49. Trudgill, P.: Sex, covert prestige and linguistic change in the urban british english of norwich. Language in Society 1(2), 179–195 (1972)
50. Wolfram, W.: The linguistic variable: Fact and fantasy. American Speech 66(1), 22–32 (1991)
51. Zhao, Q., Erdogdu, M.A., He, H.Y., Rajaraman, A., Leskovec, J.: Seismic: A self-exciting point process model for predicting tweet popularity. In: Proceedings of Knowledge Discovery and Data Mining (KDD). pp. 1513–1522 (2015)