Culture, Networks, Twitter and foursquare: testing a model of cultural conversion with social media data

Abstract

Social network research often takes the view that networks chiefly influence the spread of culture, with few reciprocal effects. While some network scholars have implied a coevolutionary relationship between the two, cultural sociologists have provided increasingly convincing evidence that it is in fact cultural preferences which mediate network structure, rather than the other way around. In the present work, we attempt to validate one such model of the conversion of cultural capital to network position. We use Twitter data to extract the ego networks of individuals and foursquare check-ins to understand their cultural preferences. Our results are indicative of the importance of considering theoretical models in which culture influences network structure. However, results are most consistent with more recent understandings of the dynamic, nuanced relationship between culture and networks.

Introduction

Our understanding of the relationship between culture and social network structure is, at best, murky. Many network scholars believe that social connections drive cultural preferences, with little reciprocal influence of culture on network structure (Pachucki and Breiger 2010). While individuals many prefer to interact with others that have similar static sociodemographic characteristics (McPherson, Smith-Lovin, and Cook 2001), the implicit claim from such scholars is that "principle influence flows *from* networks *to* culture" (Lizardo, 2006, pg. 784).

Other network researchers have come to believe that the relationship between culture and networks may be best viewed as symbiotic. As cultural preferences evolve, so to, it is argued, do our preferences for whom to interact with. This change leads to new interactions with new individuals, which in turn, provide us with another wave of new cultural preferences. This cyclic process, encapsulated in the theory of Constructuralism (Carley 1991), implies that culture and network structure *coevolve*, and thus are difficult to disentangle. While this coevolutionary perspective seems an appropriate middle ground for study, cultural sociologists have recently reinvigorated the assertion that cultural preferences are much more stable than our social connections (Lizardo 2006; Vaisey and Lizardo 2010). As cultural preferences are more stable than network relations, these scholars have suggested that cultural preferences should be considered to play the stronger causal role in the relationship between culture and social network structure.

All three of these perspectives have had an impact on the study of social media data. Viewing social structure as the causal agent in the relationship between culture and networks has helped inform studies on how cultural artifacts spread and evolve within social media data. For example, it has been shown that dynamic social relations can readily shape cultural patterns of word usage in online forums (Danescu-Niculescu-Mizil et al. 2013). Scholars who consider culture and networks as being intertwined and coevolving have given significant evidence of the nuances involved in this process (Romero, Tan, and Kleinberg 2013). However, while computational social scientists have implicitly adopted the perspective of culture as the stable force driving network structure in recent work (Quercia, Capra, and Crowcroft 2012; Chang et al. 2014), there has yet to be, to the best of our knowledge, a marriage of recent socio-theoretic efforts with this perspective to the study of social media data.

In the present work, we use a popular and relatively new theory, Omar Lizardo's *cultural conversion model* (CCM) (Lizardo 2006; 2011), that meshes the symbiotic and "culture-first" perspectives. We attempt to both validate and extend the claims of the CCM using Twitter and foursquare data. Lizardo argues that culture does not simply coevolve with network structure. Rather, the CCM suggests that individuals are constantly *using* culture in particular ways with particular social ties to obtain a particular network position. More specifically, people who have "passing knowledge" in many domains can use this *weak culture* (Schultz and Breiger 2010) to jump in at the fringes of many different social groups. In contrast, individuals who hold many varieties of *strong culture*, or deep knowledge within particular domains, can use it to form stronger bonds with like-minded individuals. The end result of this process, it is argued, is that ownership of more weak culture leads to less "closed off" (or clustered) personal networks, while ownership of more strong culture leads to more closed, tightly knit per-

sonal networks. These markers of culture are suggested to be more stable than the network connections they are used to form, and hence to play a stronger causal role in the relationship between culture and networks.

In order to perform an empirical study of the CCM, we collect data on 1,817 Twitter users who routinely post foursquare checkins as public tweets. We use these checkins as markers of the strong and weak cultural preferences these individuals hold by manually re-coding foursquare venue categories into the cultural preference domains used in a previous test of the CCM. We then extract the personal, or ego, networks of these users by crawling their tweets, follower and followee relationships and the tweets of all of the other users they have mentioned. Finally, we calculate linguistic characteristics of the user's tweet content to both relate our efforts back to previous work on Twitter and to examine the role that these lexical markers of culture may play in network structural evolution.

This vast, multi-faceted collection of data provides us with access to information on the cultural preferences, ego network structure and linguistic patterns of each user. Armed with this information, we use straightforward statistical methods to develop models for each of the following assertions of the CCM:

- The more strong cultural preferences one has, the more closed one's ego network is, while the more weak cultural preferences one has, the less closed one's ego network is
- The more total cultural preferences one has, the more total social ties one has
- The more strong cultural preferences one has, the more strong ties one has
- The more weak cultural preferences one has, the more weak ties one has

The first assertion stated is the chief theoretical contribution of the CCM. With data from the General Social Survey (GSS), Lizardo shows that individuals with more weak cultural preferences, defined by number of visits to websites in nine different high-level, topical domains, have less clustered personal discussion networks (Lizardo 2011). Similarly, individuals with more strong cultural preferences have more highly clustered personal networks. However, because he uses survey data, Lizardo's efforts are restricted to studying connections between only stronger ties in survey respondents' ego networks. In the present work, we are able to observe connections between all of a user's social ties on Twitter, thus allowing for a broader, if more noisy, study of ego network structure.

Further, while earlier empirical tests of the CCM focus on the number of weak, strong and total ties in an individual's network, these tests were carried out with a binary measure assessing the extent to which one engages in "popular" versus "highbrow" culture (Lizardo 2006). Though related to weak and strong cultural preferences, Lizardo himself notes that these previous metrics of cultural preference do not actually amount to a test of the CCM as described in his later work. Our efforts thus serve as a complement to the empirical support that currently exists for the CCM, as the lack of a formal empirical test for these latter three assertions leaves room for alternative explanations of clustering in ego networks. For example, it is possible that strong cultural preferences reduce the number of weak ties in one's network. This process would also likely lead to a reduction in ego network closure, but via an opposite process assumed by the theory.

Finally, we extend Lizardo's argument by considering his claims in the backdrop of less stable cultural forms. More specifically, we consider the CCM's model of stable cultural forms in coordination with users' lexical patterns, which prior evidence suggests are more readily subject to change. We observe an interesting connection between these stable and dynamic cultural forms we measure. This connection serves as evidence of a more nuanced model of cultural conversion, one akin to other sociological efforts (including other work by Lizardo) that admit different levels of dynamism exist in both network and cultural structures (Patterson 2014; Lizardo and Strand 2010; Vaisey and Lizardo 2010).

Related Work

A host of recent studies have considered the relationship between different forms of culture and different types of ego network structure on Twitter. Several scholars have considered the extent to which various markers of an individual's topical and cultural preferences predict the number of followers she has. Wang and Kraut (2012) use the average pairwise cosine similarity across a user's tweets as a proxy for her topical coherence, finding that "users who focus the topics of their early tweets more narrowly ultimately attract more followers with more ties among them". This claim, though echoed in earlier work (Cha et al. 2010), was refuted by later work from Hutto, Yardi, and Gilbert (2013). Hutto, Yardi, and Gilbert (2013) used the same predictor as part of a larger set of variables in a longitudinal study of Twitter follower predictors and found no evidence for the relationship between cosine similarity of a user's tweets and follower counts. Instead, the authors found that users who acted as "information providers" on Twitter, those who provided certain behavioral signals (e.g. completing their profile) and those with strong personal network structures in their follower graph had the highest number of followers.

While this line of work provides useful methodological approaches that are utilized here, it is not clear that the sociotheoretic groundings of the CCM apply to studies of follower counts. This because while the CCM focuses on social ties, there appears to be a universal agreement that unless one somehow qualifies the sociality of a tie on Twitter, observed relationships may be representative of "informational" connections rather than social ones (Ma, Sun, and Cong 2013). While an informational tie is difficult to distinguish empirically from a social tie (in part because a given relationship may involve some level of each), the social theories underlying these different types of ties are reasonably disparate. For example, it has been shown that in situations where a desire for expertise is paramount to relationship building, individuals who are perceived to span multiple social categories may be at a disadvantage in forming

these expertise-driven connections (Hannan 2010). This theoretical model directly contrasts with the assumptions of the CCM, where the ability to "code switch" between social categories is seen as an important means of increasing the number of one's (weak) social ties.

Scholars seeking to study distinctly social ties on Twitter thus have used various means to extract social relationships and ignore informational ones. The most frequent operationalizations of a *social* tie on Twitter make use of mutual following relationships, mutual retweets or mutual mentions. Though efforts have been made to calibrate better models of tie strength on Twitter (Gilbert 2012; Bak, Kim, and Oh 2012), measures of interaction frequency still seem to reliably predict relational strength in many cases in social media data (Jones et al. 2013). Additionally, because these more complex models of tie strength on Twitter rely on some of the very structural characteristics we are hoping to predict, it made more sense in the present work to rely on a frequency of interaction-based approach to measuring tie strength.

Researchers have also considered how social relationships (as opposed to informational relationships) intertwine with various forms of culture. Romero, Tan, and Kleinberg (2013) suggest that the "nicheness" of the hashtags used between two people is indicative of their tie strength, as measured by the number of mentions between them. Earlier work by similar authors also shows that different hashtags diffuse through unique network structures depending on the content they represent (Romero, Meeder, and Kleinberg 2011).

Other authors have gone beyond hashtag analyses to consider the relationship between higher-level representations of user interests. For example, Quercia, Capra, and Crowcroft (2012) use topical categorization APIs, which themselves make heavy use of named entity extraction, to categorize user tweets into broad cultural domains. The authors observe that users with higher levels of topical diversity tend to have more structural holes. While, among other differences, we here consider both depth and breadth of cultural content as opposed to diversity, such work is indicative of results presented here. Finally, Bosagh Zadeh et al. (2013) uses a classifier internal to Twitter to show that users express a relatively small number of topical interests, which helps to enforce network structure.

While none of these works specifically test the CCM, they provide us with confidence in the existence of an important, if broadly defined, relationship between cultural preferences and network structure in Twitter data. One critique that can be made of some of the works discussed in this section, however, is the implicit insinuation that lexical and higher order (e.g. topical domains) measures of culture are relatively interchangeable. Our work compliments these efforts in two ways. First, we draw on a new data source to infer the cultural preferences of Twitter users. Second, and more importantly, we make a distinction between two different forms of culture measured in previous work and consider both in a single model. We consider both lexical measures of culture, which have been shown to be relatively dynamic (Danescu-Niculescu-Mizil et al. 2013; Eisenstein et al. 2012), and culture as defined by interests in distinct topical domains, which empirical work suggests are far more stable (Lizardo 2006). We show that these two measures of culture are related, but theoretically and thus operationally distinct.

The CCM, as has been noted, focuses on stable cultural preferences. More specifically, Lizardo relies on Bourdieu's theory of practice (Bourdieu 1990), which emphasizes the study of more stable meanings of culture. In fact, practice theory suggests that linguistic measures of culture are not measures of culture at all. Bourdieu states that "language is not essential to understanding what we usually mean by culture" but rather "that most culture is implicit and exists at the levels of skills, habits, fast dispositions and implicit classificatory schemes" (pg. 9). In the present work, we retain the idea that language can in fact be construed as a weaker and less stable form of culture, if only as an artifact of the more implicit forms insinuated by practice theory. If we are interested in creating a practical bridge between Lizardo's work and data from Twitter, however, we thus require a measure of the habits and everyday actions taken by individuals. One measure of this is simply to determine the types of places that individuals go to, which we can gather from foursquare checkins.

It is worth mentioning, however, the known weaknesses in utilizing foursquare data for this purpose. Foursquare users often restrict their location sharing to locations where they wish to be seen (Lindqvist et al. 2011; Cramer, Rost, and Holmquist 2011; Tang et al. 2010). On the one hand, this process of self-presentation suggests that users will checkin at venues which they knowingly associate with a desirable image of the self, something that actually may serve as a benefit, rather than a boon, to understanding the latent cultural preferences of the user (Joseph, Carley, and Hong 2014). On the other hand, however, self-presentation may simply be a means of epitomizing a desired, rather than an actual, cultural preference.

These issues for measuring cultural preferences with foursquare check-ins extend in similar ways to using Twitter data to measure ego network structure. Well known limitations aside (Tufekci 2014; Ruths and Pfeffer 2014), the context collapse imposed by Twitter means that individuals may restrict their interactions to maintain a particular identity online that is acceptable to the audience they are trying to play to (Marwick and Boyd 2011). This collapse of context has the potential to mitigate the ability of social actors to code switch, which in turn may limit their ability to create and maintain weak ties online.

Data and Methods

We work with two primary sets of data. The first is a collection of foursquare check-ins posted publicly to Twitter, along with information on these check-ins drawn from the foursquare API. We use this data to infer the strong and weak cultural preferences of Twitter users. The second set of data is a collection of the tweets and follower and followee relationships for these users, which we use to extract ego networks and measures of the user's linguistic propensities. In this section we detail the data collection process,

how we measure strong cultural preferences, weak cultural preferences, ego network structure and linguistic traits and finally describe the statistical models we build to test our four assertions. All code and data necessary to replicate our analysis are available at **removed for blind review**.

Inferring cultural preferences

Data collection and transformation The foursquare data we use to infer cultural preferences is drawn from a merging of two different data sets of foursquare check-ins posted publicly to Twitter. The first is a data set of approximately 11M check-ins drawn by the authors of (Cranshaw et al. 2012). The second is a data set of approximately 2M additional checkins obtained from a 10% sample of all tweets sent from 2009-2012[1]. Each check-in provides us with information on the Twitter user who posted it, the time the check-in occurred and the ID of the foursquare venue the user checked in to. In total, our combined dataset consisted of slightly less than 12M check-ins (after deduplication) from around 130K users.

Via its unique ID, information on the category of each venue checked in to was obtained via the foursquare API[2]. Foursquare venue categories are hierarchical (nested to a depth of up to three), where levels are differentiated with the "::" operator. Thus, for example, the venue category "Travel Spot::Airport::Airport Terminal" refers to the venue category Airport Terminal at the third level of the hierarchy, which also falls under the "Airport" and, even more broadly, "Travel Spot" hierarchies. These venue categories are what was used to infer cultural preferences of the users, an approach taken to varying extents in several previous works utilizing foursquare data (Silva et al. 2014; Joseph, Carley, and Hong 2014; Kurashima et al. 2013).

Methods In order to extract cultural preferences from these venue categories, we decided to manually match them to the cultural forms studied by Lizardo in his empirical analysis of the CCM. While we considered taking an unsupervised approach similar to the one utilized by Chang et al. (2014) to extract domains of cultural forms from Pinterest data, we decided that a manual labeling of categories to cultural forms was likely to provide us with data more amenable to comparison with previous work. Lizardo (2011) uses nine different categories of websites to define nine types of cultural forms: (1) sports, (2) music, (3) art museum or gallery, (4) movie, (5) health, (6) game, (7) humor, (8) science, (9) a hobby.

Three human coders, two unfamiliar with the project and one of the authors, were shown 82 distinct venue categories and were asked to label them as being from one of the nine categories above, or a "none" category. Due to the fact that foursquare periodically renamed the higher levels of the category hierarchy during the course of our data collection, coders were only shown the label of the bottom of the hierarchy (e.g. "Travel Spot::Airport::Airport Terminal" was displayed as "Airport Terminal" to the coder). Pilot testing on a few categories suggested this did not appear to affect understanding of the data. We also removed restaurant, travel, residence and office categories, none of which were relevant to the cultural domains we study here, in order to decrease the burden on the coders.

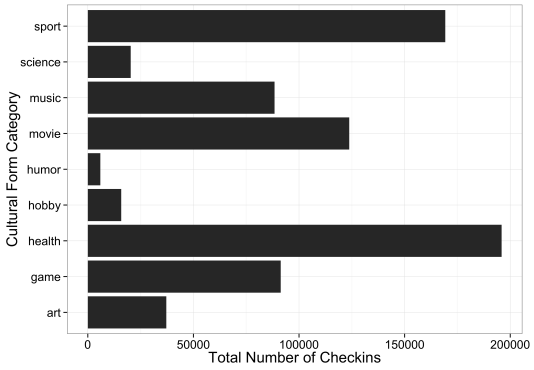

Figure 1: Number of check-ins in each of the nine different cultural domains

The two coders who were unfamiliar with the project were briefed on the categories Lizardo used, shown a paragraph from the original article where the cultural form domains were introduced and shown the original questions from the GSS that Lizardo used. Fleiss' kappa was calculated to assess the extent to which coders agreed on mappings from venue categories to cultural domains. With a value of 0.64, codings showed "substantial agreement" (Landis and Koch 1977). Figure 1 shows the total number of check-ins for each of the nine different categories[3]. As is clear, the distribution of check-ins across categories was heavily skewed. However, as it is natural to expect that certain cultural forms are more interesting to the general population than others, we choose not to control for these differences.

After obtaining the venues that were representative of each of Lizardo's cultural preference categories, we needed to determine the "strength" of the preference for each cultural form for each user in our dataset. For Lizardo, strength was absolute - respondents were classified as having a strong preference for a particular cultural form if they had visited a website relating to it three or more times, and a weak preference if they had visited one or two times. While we explored nonparametric approaches to defining preference strength in our data, we eventually decided that the most easily communicated and comparable approach was to simply mirror our definition of "strength" after Lizardo. Thus, users who had three or more check-ins in a specific cultural preference domain were deemed to have a "strong" preference for that domain. Users who had one or two check-ins in a domain had a "weak" preference for the domain.

Figure 2 shows the number of total users that had each combination of strong and weak preferences. In the figure, areas with no text represents regions where no users were present, and the intensity of the orange color is used to vi-

[1]We thank **(removed)** for this portion of the data

[2]https://developer.foursquare.com/

[3]Note the sum is considerably less than 12M, as many check-ins were to categories not related to the cultural domains of interest

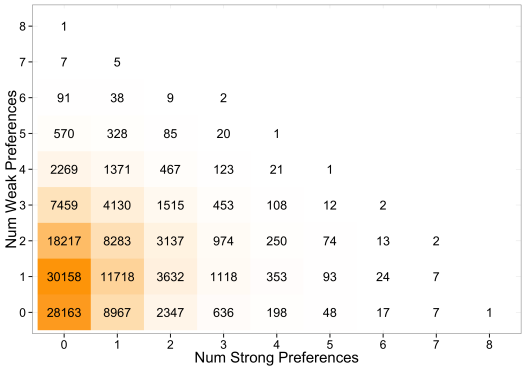

Figure 2: Number of users (defined by the grid cells in the plot) having different numbers of weak (y-axis) and strong (x-axis) cultural preferences. Orange indicates more data in the cell, and cells where no number exists are cells that no users fell into

sually depict the point that most users had between zero and two strong and weak preferences.

Twitter ego networks and linguistic content measures

Having extracted Twitter users with varying levels of strong and weak cultural preferences, we now needed to collect information about the ego network of these users and a select number of linguistic indicators. Before doing so, we first pruned our set of users in several ways. To better ensure we studied only relatively active foursquare and Twitter users, we ignore all users for whom we captured fewer than ten check-ins and who had less than 100 tweets overall. For reasons that will become clear below, we also removed users who sent less than 50 tweets in 2014. We also ignored users with greater than 25K tweets or greater than 5K followers, as they appeared to be heavily ties to promotion-based or business accounts.

As collecting the ego networks of all users that fit these conditions was computationally prohibitive, we subsampled a small collection of users to extract ego networks for. We considered only users with five or fewer strong and weak preferences, as data beyond this was judged to be too sparse. From this collection, we randomly pulled at most 1000 users from each combination of strong and weak preferences. In cases where there were less than 1000 users, we sampled all users in the cell. Due to issues in storing the set of tweets relevant to all of these users and their alters for replication purposes, we were forced to stop collection at 3175 users, randomly distributed across this larger subsample.

For each of the users of interest, we began by extracting their follower and following relationships as well as the maximum number of their tweets allowed by the REST API (3200). If a user in our sample had protected their account since the collection of our check-in dataset, we dropped the user from our sample. After obtaining a particular user's tweets, we drew out all of her social ties. Our definition of a social tie is slightly more restricted than previous work, as we combine two of the traditional metrics of social connections on Twitter. Consider a potential undirected social tie $< i, j >$, where i and j are two Twitter users. Additionally, consider the value $f(i, j)$, which is a boolean indicating whether or not i follows j, and the value $n_m(i, j)$, which is the number of times that i has mentioned j *in the past year*, where we use the 2014 calendar year as our year of study. This temporal restriction admits the volatility in social relationships, and is used in the GSS to determine social connections. We consider a social tie to exist iff $f(i, j)\,\&f(j, i)\,\&min(n_m(i, j), n_m(j, i)) > 0$. The use of the min function ensures that this tie is reciprocal, an important feature of social ties both online and off. A tie is strong if $min(n_m(i, j), n_m(j, i)) > 2$. A tie is weak if $< i, j >$ meets the conditions for a social tie but does not meet this strong tie condition[4].

After all restrictions had been put in place, we had a final N of 1817 users. For each of these users, we completed our network data extraction from Twitter by collecting the following and follower relationships for each of the user's social ties, along with the tweets of each of these actors. A first-order ego network was then constructed by adding social ties between a user's alters where the alters' relationship fit the definition of a social tie described above.

After drawing out all network information for these users, we then extract three linguistic markers of a user's tweets that have been utilized in prior studies. First, we extract the proportion of a user's tweets expected to contain informational content, as defined by Hutto, Yardi, and Gilbert (2013). Second, as used in Hutto, Yardi, and Gilbert (2013) as well, we determine the average number of hashtags per tweet for each user. Finally, as used by both Hutto, Yardi, and Gilbert (2013) and Wang and Kraut (2012), we extract the average pairwise cosine similarity in a user's tweets. Our tokenization approach uses the base tokenizer from O'Connor, Krieger, and Ahn (2010) and appends several post-processing steps, including the removal of stopwords[5]. Previous articles do not report their procedure, so our assumption is simply that they engaged in a similar approach.

These linguistic markers are extracted from only the users' tweets sent before 2014, and thus precede tweets used to construct the users' ego networks. This is also true of the data used to extract users' cultural preferences, as data collection for the foursquare check-ins ended in 2012, a full two years before the network data we collect. This temporal ordering allows us to discuss the effect of a user's prior linguistic behavior and cultural preferences on their actions in a future time period. As we will see, however, while these measures are calculated on data that occurred prior to network extraction, the reciprocal relationship of culture and networks has important implications on our understanding of results.

[4]Note that we tried various realistic setting for this parameter, none of which appeared to have an effect on the qualitative findings we report here

[5]Tokenization code is included in the code release for this article

Regression Models

We present results for four statistical models, one for each assertion we derive from the CCM. We fit one model each to measure the effect of total, strong and weak cultural preferences on the numbers of total, strong and weak ties, respectively. Additionally, we fit a model for network closure, where the dependent variable is, as Lizardo uses, the number of social ties between a user's alters. In this model, we exclude users in our dataset with one or fewer social ties, as it is impossible for there to be a connection between alters in this case (N=1459).

In determining which family of statistical modeling was appropriate for our data, we considered a variety of approaches. Our considerations included both linear and non-linear models, various families in the generalized linear modeling framework and models with and without zero-inflation and regularization. After considering the available options, we determined that negative binomial regressions showed acceptable fits to the data across all four dependent variables while remaining understandable to an audience familiar with only more traditional statistical approaches.

Because we study four different questions, we found it practical to use a Bonferroni correction on the typical $\alpha = .05$ to ensure that we controlled for the large number of comparisons we made. Thus, $\alpha = .01$ was used to determine significance. All coefficients discussed in the following section are significant with $p < .01$. Additionally, the models presented are parsimonious, as determined by starting with the full predictive model described below and then selectively excluding uninteresting variables using ANOVAs (again with $\alpha = .01$) to compare the nested models. All models show a reliable ($p < .01$) fit to the data. The manual model selection process used allowed us to visually assess the fit and practicality of each successively simpler model given our theoretical backdrop and was thus preferable to automated model selection approaches, some of which have additional methodological issues as well (Derksen and Keselman 1992).

Before model selection, all four models include our three linguistic variables derived from users' pre-2014 tweets. We also include in all models three other variables to control for the level of activity of a user. We use the logarithm of the number of check-ins a user has to control for foursquare activity, as well two additional logged variables to control for Twitter activity, one each for the total number of mentions and total number of tweets a user had in 2014. In cases where the CCM predicts an effect of strong and/or weak cultural preferences, we include both as predictors in the initial, full model. In the single case where the CCM predicts an effect of the total number of cultural preferences instead of a unique effect of strong vs. weak culture, we use total preference counts as opposed to including both strong and weak counts as predictors. Finally, in the closure model, we follow Lizardo (2011) and include an offset term in the regression model for the logarithm of the total number of possible connections (i.e. the number of ties squared). While it is possible that this variable might "over-control" for the possible number of ties (Anderson, Butts, and Carley 1999), various models we fit with and without the offset reached similar qualitative conclusions.

All coefficients in all models are standardized by subtracting the mean and dividing by two standard deviations, an approach that allows for an easier comparison of low versus high values on the independent variables' intrinsic scales (Gelman 2008)[6]. This standardization is also useful in that it allows a direct comparison across variables within a single regression model. Finally, as opposed to considering model coefficients, we here follow Wang and Kraut (2012) and display the *Incident Rate Ratio* (IRR) of the outcome variables. The IRR can be interpreted as a multiplicative effect that a two standard deviation change in the independent variable has on the dependent variable. We discuss these effects in terms of percentages in the following section, and use brackets [] to provide 95% confidence intervals (CIs) for parameter estimates.

Results

Figure 3 displays coefficients, excluding the intercept, for the most parsimonious models for predicting, from left to right, the number of total, strong and weak social ties of the Twitter users we study. Unsurprisingly, the logarithm of the number of a user's total mentions in 2014 is the strongest predictor for each type of tie. Beyond this obvious conclusion, Figure 3 also shows that the total tie and weak tie models provide support for two of the assertions we posed regarding the CCM. A two standard deviation increase in a users' total number of cultural preferences is associated with an 18.6% [7.8-30.4%] increase in the users' total number of social ties. Similarly, users with high levels of weak cultural preferences have, on average, almost 14% [4.6-22.9%] more weak ties than those with low levels of cultural preferences[7]. As the middle plot in Figure 3 shows, however, there is no significant effect of strong cultural preferences on the number of strong ties that a user has.

Figure 3 also shows that the only other variable appearing in each of the tie count models is the pairwise cosine similarity of a user's tweets prior to 2014. This variable is negatively associated with the number of strong, weak and henceforth total number of social ties for a user. In fact, it is the most heavily negative predictor in all three cases, and is the second strongest predictor in absolute magnitude beyond the mention count control. Users with low levels of cosine similarity in their tweets prior to 2014 have, on average, around only 65% of the strong, weak and total ties that users with higher levels of linguistic similarity do.

The only other variables we observed that were negative predictors of tie count were proportion of tweets containing informative content, which had a negative effect on total (15.1-22.0% decrease) and strong (36.0-48.4% decrease) tie counts, and the number of check-ins a user had, which had a

[6]While this provides a useful interpretation of the results, we note that because we compare across two standard deviations and thus across large disparities in the independent variables, the reported effect sizes are high

[7]Note this effect is virtually the same (mean effect of 12%) if we kept the (unreliable) logarithm of checkin counts in the model as a control

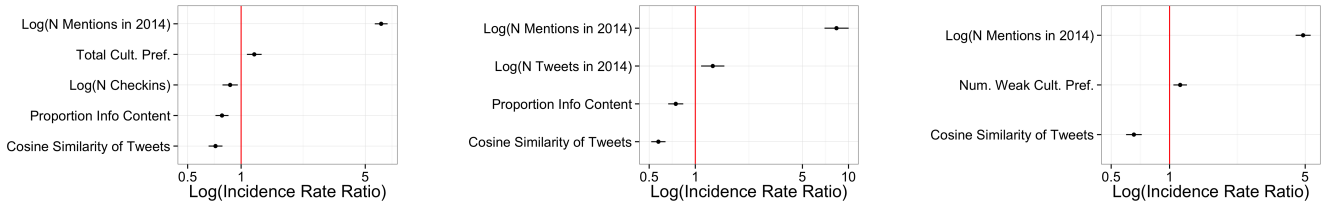

Figure 3: Regression coefficients with 95% CIs for the three tie count models. From left to right, we display the total tie count model, the strong tie count model and, finally, the weak tie count model results. The red line at an IRR of 1 indicates the value at which the independent variable would have no effect on the dependent variable.

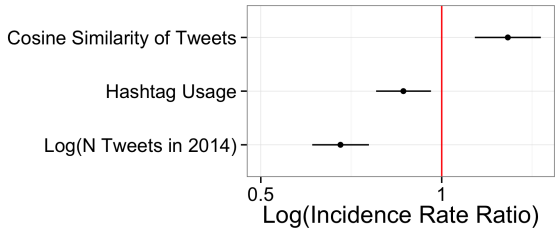

Figure 4: Regression coefficients with 95% CIs from the tie closure model. The red line at an IRR of 1 indicates the value at which the independent variable would have no effect on the dependent variable.

weaker but reliable negative effect on the total number of ties an actor had (7.0-15.8% decrease). This latter observation is anecdotally interesting in that it seems to confirm user fears that too many check-ins can annoy others (Cramer, Rost, and Holmquist 2011), leading to a reduction in social connections.

Figure 4 plots coefficients for the final model we test, the network closure model. Due to the use of the offset variable regarding the total number of possible ties amongst users' alters, all coefficients are here interpreted relative to the possible number of connections between their social ties. We find no support for the claims of the CCM in our data, as neither strong nor weak cultural preferences emerge as significant predictors of network closure. The only predictors to remain in the parsimonious model of network closure are average hashtag usage, the number of tweets a user sent in 2014 and the lexical coherence of a user's tweets prior to 2014 as measured via cosine similarity. Heavy users of hashtags have, on average, only 86% [77.8-95.6%] as many connections among their alters as users who infrequently use hashtags. Users who are more active have, on average, only 67.8% [60.8-75.6%] of the number of ties amongst alters than those who are less active. In contrast, users who have more linguistic similarities across their tweets prior to 2014 have, on average, 29% [13.5-46.1%] more ties amongst alters in their ego network.

Discussion

In this section, we summarize model results and place them within the context of a) the "socialness" of Twitter and b) the relationship between culture and networks.

Social and informational ties on Twitter

Our results further imply the importance of considering Twitter as both a *social* and a *media* website (Kwak et al. 2010). Most obviously, we find that users who are interested in engaging in the spread of information on Twitter are likely to do so at the expense of prolific interaction with any particular subset of their social connections. It is also quite possible that the negative effect of sending more tweets on network closure is related to the extent to which a user engages in information-spreading behaviors over social behaviors on Twitter. Although variance inflation factors, among other diagnostics, implied the two variables explain unique variances in our statistical models, there is a moderate, positive correlation between the log of the number of tweets a user sent in 2014 and the information content in their previous tweets ($\rho = .37$, 95% adj. bootstrap percentile CI [0.33, 0.41] with 10,000 iterations). Why one variable is more predictive of tie counts while the other reliably predicts closure is unclear, but future work may use a combination of these variables to identify actors early on who will engage in information spreading behavior over relational maintenance with strong ties in the future.

This result speaks to the tension between the social and informational aspects of Twitter. On the other hand, we also find further evidence that the combination of the social and informational mechanisms available on Twitter can have reinforcing effects. Yang et al.'s (2012) work on the dual natural of hashtags shows that hashtags are used both as markers of distinct content and as markers of entrance into a particular community. This helps to explain why hashtag usage is negatively associated with network closure, as the use of hashtags as both a vehicle for information spread and as a mechanism to gain entrance into communities would both insinuate an association between increased use of hashtags and less closed social networks.

In sum, while our efforts in the present work were geared towards understanding social relationships on Twitter, we found great value in utilizing variables developed in previous work that help to understand the contrasting and reinforcing effects of the informational nature of Twitter on user's social network structure. Perhaps most interestingly, our results surrounding these variables suggest users exist on a spectrum between those who engage in relational maintenance with strong ties and those who focus on information when using the site.

Unraveling the relation between culture and networks

Relevant to the CCM, we find that weak cultural preferences determined using data from 2012 have a reliable effect on Twitter ego networks constructed from tweets sent *12-24 months* later. Our work thus adds novel empirical evidence to the increasingly popular view that culture has a stable and profound effect on network structure. However, these findings must be qualified in two important ways. First, strong cultural preferences have no effect on strong social ties, nor on ego network closure. It is distinctly possible that the "weak tie" nature of Twitter (Gilbert 2012; Hutto, Yardi, and Gilbert 2013) precludes the study of the true impact of strong cultural preferences on network structure. Additionally, there is the possibility that Lizardo's empirical focus only on connections amongst strong ties (due to his use of GSS data) means that his empirical findings do not hold for larger networks. However, a re-analysis of our data finds that strong cultural preferences do not predict closure in ego networks consisting of only strong ties, thus leading us to prefer the former explanation.

The second caveat to our findings related to the CCM is that weak cultural preferences, while predicting an increase in the number of a user's weak ties, do not decrease closure in a user's ego network as we expected. Instead, it is the cosine similarity of users' tweets that has the expected negative association with network closure. Cosine similarity of a users' tweets also predicts a strong decrease in their number of social ties. Neither of these findings can be remedied by the theoretical guidelines established by the CCM, as language is well known to vary dynamically via social influence (perhaps particularly on Twitter; Eisenstein et al. 2012).

Both of these findings are, however, consistent with the coevolutionary perspective on networks and culture, which admits the instability of both and their intertwinement. More specifically, Constructuralism (Carley 1991) predicts that an actor with a more restricted set of "knowledge", here operationalized as words in her vocabulary, should have both a smaller and more closed social network. Actors with restricted vocabularies should have smaller social networks due to the fact that they are only interested in the small subset of others who share the same refined set of terms they use, and more closed ego networks because their alters should thus also share more restricted terminology amongst themselves. This interpretation of results, supported in part by Romero, Tan, and Kleinberg's (2013) findings as well, seems appropriate given that Twitter is heavily based on text. Users are thus frequently forced to use textual cues as an important piece in the larger set of signals (Hutto, Yardi, and Gilbert 2013) available to them to determine whom to form and maintain social interactions with on the site.

Under a Constructuralist interpretation of our results, however, the fact that our linguistic similarity measure is taken on tweets that occur before those used to construct the network of interest becomes unimportant. This is because these linguistic similarities can simply be construed as the byproduct of previous network structures similar to those we observe later on. There thus exists not causal claim

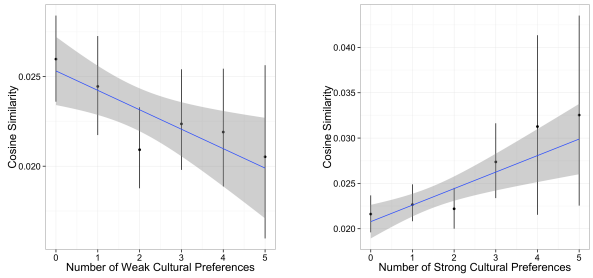

Figure 5: On the left, a plot of the relationship between weak cultural preferences and the cosine similarity of a user's pre-2014 tweets. Black vertical bars are 95% bootstrapped CIs (with a point at the mean) of cosine similarity at each weak tie count. The blue line is a best-fit least-squares estimate of the linear relationship between the variables, with its own 95% CI in grey. The same plot, except in comparison to strong cultural preferences, is displayed on the right

to be made. However, there does exist a causal story that posits some stable, external propensity of an actor to have a high level of consistency in their language, which in turn may lead to smaller, more clustered personal network.

Though this is not the exact stable propensity implied by the CCM, this interpretation is consistent with a slightly more generic cultural conversion model in which stable cultural schemata influence the emission of more dynamic cultural artifacts, which in turn co-evolve with network structure. If this were to be the case, we would thus expect that an increase in weak culture is associated with less linguistic similarity in users' tweets, while more strong cultural preferences are indicative of more consistent language. Figure 5 shows, on the left, a negative, significant ($p < .001$) association between lexical coherence and the number of weak cultural preferences one has. On the right, we observe a positive, significant ($p < .001$) association between lexical coherence and the number of strong cultural preferences one has. This relationship is not at all obvious or pre-ordained. Indeed, the check-ins used to determine cultural preferences are frequently not even included in the set of user tweets available from the API that we used to calculate linguistic coherence. Our data thus support the idea that stable cultural preferences influence less stable linguistic markers of a user's cultural embeddings, which in turn exist within a symbiotic relationship with network structures.

Conclusion

The present work is motivated by the ongoing debate over the relationship between culture and networks (Pachucki and Breiger 2010). We extract a measure of cultural preferences from foursquare data and ego networks from Twitter in an attempt to better understand one particular model of the conversion from cultural preferences to network structure. In doing so, we make three contributions to the study of social media data, culture and networks.

First, we provide a novel combination of data from multiple social media websites. This effort complements recent work combining data from a variety of social media sites in

order to better understand user behavior (e.g. Tang, Lou, and Kleinberg 2012). Second, we provide a novel take on the way in which users exist along a quantifiable continuum from social to information usage patterns on Twitter and how this may affect the structure of their ego networks. This complements the existent literature on the dual role of Twitter as a social and informational network (Kwak et al. 2010; Yang et al. 2012). Finally, and most importantly, we provide a novel study of the interplay between culture and networks using social media data. This work complements recent work on the relationship between user's topical preferences and the structure of their social networks online (Quercia, Capra, and Crowcroft 2012; Wang and Kraut 2012; Hutto, Yardi, and Gilbert 2013; Chang et al. 2014).

As with any study that uses social media data, myriad methodological issues may have hindered or played a mediating role in our results (Tufekci 2014; Ruths and Pfeffer 2014). There are, however,three issues specific to our efforts. First, it is unclear how well foursquare checkins, or the way in which we divided them into strong and weak preferences, really detail the true cultural preferences of users. While we feel the use of check-in data comes at least as close to the definition of cultural preferences provided by the CCM as survey data Lizardo himself used, it is possible that check-ins still fail to capture the desired information. Future work should address this by considering how multiple, distinct markers of stable cultural preferences relate. Second, foursquare users are a very particular subset of Twitter users, and our results may not necessarily generalize beyond them. Finally, it is unclear whether or not our notions of social ties, be it strong, weak or in general, are really measuring a user's "true" social ties. While network surveys are themselves notoriously poor at capturing an individual's true social network, Gilbert's (2012) work suggests there is a way to combine survey and social media data in a way that provides a better model of ego networks than either alone can provide.

Limitations aside, our work provides interesting empirical insight into the ongoing debate over the relationship between culture and networks. Our work suggests the enticing possibility that, as is so often the case, everyone is right. Our results are consistent with a world in which there are certain elements of culture that are highly stable and thus cannot be readily changed via social interaction. These stable cultural forms may have strong effects, in part through less stable cultural artifacts, on the structure of our evolving social networks. Our findings do not preclude the existence, however, of strong social ties which are themselves unaffected by cultural preferences, thus forming a backbone of sociality that deeply affects less stable cultural preferences. Finally, the exchange of the ephemeral elements of culture and the transitory nature of social ties combine to form a mezzo-level, symbiotic linkage between culture and network forms. In such a model, both cultural and network structures exist on a spectrum of dynamism, where more dynamic network elements are more amenable to mediation by more stable cultural elements as well as the other way around.

This depiction of culture and networks falls in the spirit, if not in the precise assumptions as they are understood here, of Lizardo's cultural conversion model (Lizardo 2006; 2011). It has also been implied in several other recent discussions of the interplay of culture, cognition and networks (Lizardo and Strand 2010; Patterson 2014). In this way, our work thus serves as yet another example of how the right combination of insight from social theory and the richness of data that social media offers can provide meaningful insight into contemporary questions of sociological interest. We look forward to additional studies by sociologists, both computational and not, on different approaches to measuring higher-order cultural structures (e.g. Yang et al. 2014) and the "dynamically stable" (Patterson 2014, pg. 22) nature of its relationship with network structure .

References

Anderson, B. S.; Butts, C.; and Carley, K. 1999. The interaction of size and density with graph-level indices. *Social Networks* 21(3):239 – 267.

Bak, J. Y.; Kim, S.; and Oh, A. 2012. Self-disclosure and relationship strength in twitter conversations. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL '12, 60–64. Stroudsburg, PA, USA: Association for Computational Linguistics.

Bosagh Zadeh, R.; Goel, A.; Munagala, K.; and Sharma, A. 2013. On the precision of social and information networks. In *Proceedings of the first ACM conference on Online social networks*, 63–74.

Bourdieu, P. 1990. *The logic of practice*. Stanford University Press.

Carley, K. 1991. A theory of group stability. *American Sociological Review* 56(3):331–354. ArticleType: research-article / Full publication date: Jun., 1991 / Copyright 1991 American Sociological Association.

Cha, M.; Haddadi, H.; Benevenuto, F.; and Gummadi, P. K. 2010. Measuring user influence in twitter: The million follower fallacy. *ICWSM* 10:10–17.

Chang, S.; Kumar, V.; Gilbert, E.; and Terveen, L. 2014. Specialization, homophily, and gender in a social curation site: Findings from pinterest.

Cramer, H.; Rost, M.; and Holmquist, L. E. 2011. Performing a check-in: emerging practices, norms and 'conflicts' in location-sharing using foursquare. In *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, MobileHCI '11, 57–66. New York, NY, USA: ACM.

Cranshaw, J.; Schwartz, R.; Hong, J. I.; and Sadeh, N. 2012. The livehoods project: Utilizing social media to understand the dynamics of a city. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media*, ICWSM '12. AAAI.

Danescu-Niculescu-Mizil, C.; West, R.; Jurafsky, D.; Leskovec, J.; and Potts, C. 2013. No country for old members: User lifecycle and linguistic change in online communities. *WWW13*.

Derksen, S., and Keselman, H. J. 1992. Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology* 45(2):265–282.

Eisenstein, J.; O'Connor, B.; Smith, N. A.; and Xing, E. P. 2012. Mapping the geographical diffusion of new words. *arXiv:1210.5268*.

Gelman, A. 2008. Scaling regression inputs by dividing by two standard deviations. *Statistics in medicine* 27(15):2865–2873.

Gilbert, E. 2012. Predicting tie strength in a new medium. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, CSCW '12, 1047–1056. New York, NY, USA: ACM.

Hannan, M. T. 2010. Partiality of memberships in categories and audiences. *Annual Review of Sociology* 36(1):159–181.

Hutto, C. J.; Yardi, S.; and Gilbert, E. 2013. A longitudinal study of follow predictors on twitter. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 821–830.

Jones, J. J.; Settle, J. E.; Bond, R. M.; Fariss, C. J.; Marlow, C.; and Fowler, J. H. 2013. Inferring tie strength from online directed behavior. *PLoS ONE* 8(1):e52168.

Joseph, K.; Carley, K. M.; and Hong, J. I. 2014. Check-ins in "blau space": Applying blau's macrosociological theory to foursquare check-ins from new york city. *ACM Trans. Intell. Syst. Technol.* 5(3):46:1–46:22.

Kurashima, T.; Iwata, T.; Hoshide, T.; Takaya, N.; and Fujimura, K. 2013. Geo topic model: joint modeling of user's activity area and interests for location recommendation. In *Proceedings of the sixth ACM international conference on Web search and data mining*, WSDM '13, 375–384. New York, NY, USA: ACM.

Kwak, H.; Lee, C.; Park, H.; and Moon, S. 2010. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, WWW '10, 591–600. New York, NY, USA: ACM.

Landis, J. R., and Koch, G. G. 1977. The measurement of observer agreement for categorical data. *biometrics* 159–174.

Lindqvist, J.; Cranshaw, J.; Wiese, J.; Hong, J.; and Zimmerman, J. 2011. I'm the mayor of my house: examining why people use foursquare - a social-driven location sharing application. In *Proceedings of the 2011 annual conference on Human factors in computing systems*, CHI '11, 2409–2418. Vancouver, BC, Canada: ACM.

Lizardo, O., and Strand, M. 2010. Skills, toolkits, contexts and institutions: Clarifying the relationship between different approaches to cognition in cultural sociology. *Poetics* 38(2):205–228.

Lizardo, O. 2006. How cultural tastes shape personal networks. *American Sociological Review* 71(5):778–807.

Lizardo, O. 2011. Cultural correlates of ego-network closure. *Sociological Perspectives* 54(3):479–487.

Ma, Z.; Sun, A.; and Cong, G. 2013. On predicting the popularity of newly emerging hashtags in twitter. *Journal of the American Society for Information Science and Technology* 64(7):1399–1410.

Marwick, A. E., and Boyd, D. 2011. I tweet honestly, i tweet passionately: Twitter users, context collapse, and the imagined audience. *New Media & Society* 13(1):114–133.

McPherson, M.; Smith-Lovin, L.; and Cook, J. 2001. Birds of a feather: Homophily in social networks. *Annual Review of Sociology* (1):415–444.

O'Connor, B.; Krieger, M.; and Ahn, D. 2010. TweetMotif: Exploratory search and topic summarization for twitter. In *ICWSM*.

Pachucki, M. A., and Breiger, R. L. 2010. Cultural holes: Beyond relationality in social networks and culture. *Annual Review of Sociology* 36(1):205–224.

Patterson, O. 2014. Making sense of culture. *Annual Review of Sociology* (0).

Quercia, D.; Capra, L.; and Crowcroft, J. 2012. The social world of twitter: Topics, geography, and emotions. In *ICWSM*.

Romero, D. M.; Meeder, B.; and Kleinberg, J. 2011. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proceedings of the 20th international conference on World wide web*, 695–704.

Romero, D. M.; Tan, C.; and Kleinberg, J. 2013. On the interplay between social and topical structure. In *Proc. 7th International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Ruths, D., and Pfeffer, J. 2014. Social media for large studies of behavior. *Science* 346(6213):1063–1064.

Schultz, J., and Breiger, R. L. 2010. The strength of weak culture. *Poetics* 38(6):610–624.

Silva, T. H.; de Melo, P. O.; Almeida, J.; Musolesi, M.; and Loureiro, A. 2014. You are what you eat (and drink): Identifying cultural boundaries by analyzing food & drink habits in foursquare. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, ICWSM '14. AAAI.

Tang, K. P.; Lin, J.; Hong, J. I.; Siewiorek, D. P.; and Sadeh, N. 2010. Rethinking location sharing: exploring the implications of social-driven vs. purpose-driven location sharing. In *Proceedings of the 12th ACM international conference on Ubiquitous computing*, Ubicomp '10, 85–94. New York, NY, USA: ACM.

Tang, J.; Lou, T.; and Kleinberg, J. 2012. Inferring social ties across heterogenous networks. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, 743–752. New York, NY, USA: ACM.

Tufekci, Z. 2014. Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *ICWSM 14: Proceedings of the 8th International AAAI Conference on Weblogs and Social Media.*

Vaisey, S., and Lizardo, O. 2010. Can cultural worldviews influence network composition? *Social Forces* 88(4):1595–1618.

Wang, Y.-C., and Kraut, R. 2012. Twitter and the development of an audience: those who stay on topic thrive! In *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems*, 1515–1518.

Yang, L.; Sun, T.; Zhang, M.; and Mei, Q. 2012. We know what @you #tag: does the dual role affect hashtag adoption? In *Proceedings of the 21st international conference on World Wide Web*, WWW '12, 261–270. New York, NY, USA: ACM.

Yang, S.-H.; Kolcz, A.; Schlaikjer, A.; and Gupta, P. 2014. Large-scale high-precision topic modeling on twitter. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 1907–1916. ACM.