Characterizing Semantic Change with Interactive Visualization: A Case Study in Atlanta Online Communities

Yuebin Dong, Hang Jiang, Diego Lestani, Amanda Horne

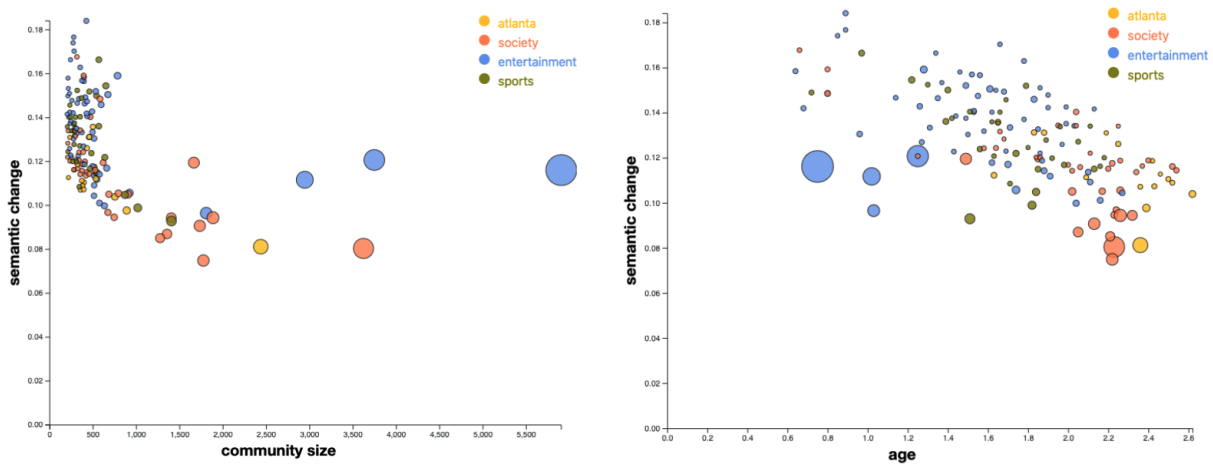

Fig. 1. **Semantic change & social factors** Relationship between semantic change (y-axis) and community-level social factors (x-axis).

1 Introduction

Language on the Internet is often used in non-standard and innovative ways [1]. Many works in sociolinguistics have shown that social variables (e.g., ethnicity, age, gender, social status) are closely related to linguistic variation in online communities [10, 16, 25]. These communities tend to develop their sociolects, or social dialects, which is similar to the concept of regional dialects [25]. Such linguistic variation includes orthographic variants [34] as well as reductive/innovative meaning change (emergence/loss of a full-fledged meaning of a word) [3, 33]. Therefore, one user's psychological encoding of a word also changes according to the community-specific context. For example, the connotative meaning of "police" or "school" may vary significantly for a speaker depending on their race, socioeconomic class, and geographical context (e.g, urban vs. rural). In this work, we hope to understand how "dialects" on social media vary from the norm across communities.

We focus our study on the city of Atlanta for several reasons. As background, the work is involved in a collaborative project on the ground in Atlanta to leverage local listening and understanding of communities to inform public health communications. Thus, studying online communities based in Atlanta complements this ongoing work. Furthermore, although online interest communities need not have any grounding in geographic locale, location is a factor in defining linguistic communities. Atlanta, as a major metropolitan center of the Southeast, contains rich linguistic variation and change [29, 32]. In the past, researchers [19, 29] have studied linguistic variation in Atlanta with surveys and interviews, but none to our knowledge have used abundant social media data to characterize the linguistic variation in the city. The central goal of the paper is to understand the relation of linguistic variation and social factors in a geographically rooted collection of communities on social media with an interactive data visualization tool.

In particular, we use the `TwitterATL` dataset for our study, which contains 149 Atlanta interest-based communities on Twitter with the community-level linguistic variation and social factors. In this project, we develop three interactive data visualizations to explore patterns of linguistic variation in Atlanta communities.

This project is partially based on the project from one teammate's project in Center for Constructive Communication, MIT Media Lab.

2 Related Work

Linguistic Variation on Social Media. The growth of social media platforms has allowed sociolinguists to conduct large-scale studies in language variation [25], which is not practical using traditional methods such as surveys and interviews. Many related works have focused on how lexical variation from the norm is related to social and linguistic factors [6, 11, 27, 35]. The community-specific linguistic norms and differences are often distinguished by which words are used. Recently, [39] proposed a PMI-based approach to identify the community-specific language use of Reddit communities. Later, [22] proposed a BERT-based approach to encode semantic variation on 474 Reddit communities. However, there are far fewer works that utilize state-of-art techniques to study community-specific language variation. [2] first proposes dialect-aware skip-grams to study regional dialects and [9] applies the method on Reddit sociolects. To extend the previous works to a larger set of communities, [22] employs BERT to characterize English variation in 474 Reddit communities, demonstrating promising directions to study community-specific language variation with pre-trained language models. Finally, [37] proposes social attention to overcome language variation on Twitter for sentiment analysis, showing that attention-based approaches can be used to address language variation on social networks.

Data Visualization in Language Change. Visualizations in language variation studies are typically static developed by Python or R, thus unable to provide users with enough information and insights through flexible interaction. For instance, [18] develop both static bar charts and line plots to show language variation over time. This work also visualize the semantic change of words by annotating the word with nearest neighbors at different times with an arrow plot. [9, 22] both use scatter plots to show the relationship between two factors and use horizontal bar charts to analyze the semantic changes at a

category level. None of these works have provided interactive design to study linguistic variation. Inspired by the observable demos in D3 library[1] [5], we decide to develop an interactive framework in D3 to analyze linguistic change.

Overall, there are **two major contributions** of this work. First, we design an interactive website for researchers to study linguistic variation in online communities. Second, we conduct analysis on the Atlanta Twitter to explore the relationship between semantic changes and social factors in both social identity and interaction. We observe similar patterns between linguistic variation and community attributes (e.g., community size, user activity) on Twitter as previous works found on Reddit [9, 22]. In this study, we introduce additional social factors such as social identity (age, gender, organization status) from Twitter profiles for analysis and have interesting findings.

3 METHODS

3.1 The Dataset

The dataset for our project was provided by Hang Jiang from his work at the MIT Media Lab[2]. Because we had a different goal with this data - helping users visualize it - we made a few modifications to the dataset. Our first step in modifying the data was to clean up the different columns and only keep the columns that we thought the users would be interested in. For each cluster, we kept the identifiers, cluster names, semantic change, and twitter information related to that cluster like the top 25 accounts, stats about friend count and follower count, typical topic words found on their tweets, and stats about twitter activity like tweeting and favoriting. We also kept demographic information like age, gender ratio and community size. Overall, there are 149 communities and the following key (not comprehensive) attributes (columns):

- **Semantic Change** is the average semantic shift of 1000 frequent terms between the selected community and a norm community[3].
- **Age** is encoded four age groups $<= 18$, 19-29, 30-39, $>= 40$ into 0-3 and use the average age of community members to represent the community-level age.
- **Gender Diversity** is measured by a community's female ratio.
- **Organization Status** is represented the proportion of organizational accounts in each community.
- **Community Size** is the number of unique users in a
- **Social Status** is measured by the median number of friends per user in a Twitter community
- **User Activity in Tweeting** is the average number of tweets per user in the community.
- **User Activity in Favoriting** is measured by the average number of favorites given out per user in a community.
- **Closeness** of a community is the average closeness score per user. This closeness score of a user measures its average inverse distance to all other users.
- **Betweenness** of a community with the average betweenness score per user. It represents the degree to which nodes stand between each other.
- **Category** of a community is characterized by the topics of a community and there are four major categories (entertainment, sports, atlanta, society).

Once we properly cut down the dataset to the key information we needed, we then assigned each cluster to a cluster group based on the typical content/topic of their twitter activity. We then also took these cluster groups and assigned them to a broader category. One example of this was taking the original cluster about International Soccer, assigning it to the cluster group "Soccer" since there was multiple clusters about soccer, and then assigning it to the category "Sports". This ended up being useful in one of our visualizations that allowed users to explore the data in different layers with these 3 different parts of a cluster: the category, the cluster group and the cluster. Once complete, we then had our dataset that we'd use throughout our visualization. All of these modifications were done without sacrificing the integrity or losing any of the information that was provided in the original dataset from the MIT Media Lab.

3.2 Visualization Techniques

When implementing the visualization, we used many techniques that we had learned throughout the semester. Throughout our website there were many places where we filtered the data based on what exploration option user chose. One example of this is the user clicking different parts of the circle graph. To start out, the user sees the broad level category names of "Atlanta", "Sports", "Entertainment" and "Society". Then when they click on one of them, they see the different cluster groups within that category. Then when they click on a cluster group, it shows them individual clusters in that cluster group, and finally when they click on an individual cluster it shows them the top twitter follower usernames. This technique of filtering proved to be useful in making it simpler for the user to explore the data at both a broad and deep level, while not making it too complicated or overwhelming in the beginning.

Additionally, based on peer review feedback, we used the first part of the website to help users understand what semantic change meant, and pushed our other visualizations to the bottom for the best chances that the user would understand what semantic change meant by the time they got to that part of the website.

3.3 Website Implementation

To program the visualization on the website, we used D3 because we were familiar with that from former assignments in the semester. D3 also was useful because of the many different built in functions that were available to help us display our data in a certain way. We also used Vue for our website to make it look nicer because it offered a better layout and aesthetic look, and it was compatible with our D3 code. For nearly all of the different parts of our visualization we utilized D3 algorithms whenever we needed to sort our data or perform computation on it such as averaging.

4 RESULTS

Our system produces a set of visualizations to understand the problem of semantic change in Atlanta's online communities. This understanding comes from defining "communities" as the unit of observation and grouping categories to help the user digest the information sequentially. As mentioned before, the chosen visual encoding is a packed set of circles which can be resized according to a set of metrics. This sequential approach in the visualization is composed of three steps.

4.1 Step 1. Understanding community clusters

The first one (Figure 2, 3, and 4) is a bubble chart that provides the user with a macro-micro approach that goes for from four broad categories through a cluster that serves as an intermediate layer of grouping but with higher detail than the previous broad categories, and a third layer where the maximum level of granularity is achieved: community. In addition, there is a fourth layer that has not to do with the grouping hierarchy but that shows the main twitter accounts that the selected community follows. The first layer where the broad categories can be seen will let the user know, for example, that the largest communities in terms of size are not necessarily the ones with the largest follower count.

4.2 Step 2. Ranking communities

After this first immersion in Atlanta's online communities and it's following pattern, the user will be able to explore rankings of those communities based on a set of metrics (Figure 5). The scrolling and zooming tool allows the user to focus in as many categories as they want and not lose sight of the rankings due to the large number of

[1] https://observablehq.com/@didoesdigital/21-june-2020-brushing-to-filter-and-zoom-using-d3-brush, https://observablehq.com/@d3/focus-context, https://bl.ocks.org/bumbeishvili/6c54d3f0e202aa7004a669a768369c5d

[2] Details for data collection is in the appendix A.

[3] Details about the semantic change calculation is in the Appendix B.

communities. This functionality will let the user know, for instance, that while "follow", "sex" and "Trump" are the largest communities in size, those with highest number of followers are: "Basketball", "care" and "entrepreneur".

4.3 Step 3. Visualizing relationships

The third step in the sequence allows the user to understand relationships between variables (Figure 4.4). This is done through a scatter plot where the size of the dots represent the community size and their color represents one of the four broad categories. Examples of insights that can be gained with this visualization are: a) detecting trends for different pair of factos, b) trends specifics to the overall population or to a specific category (through filtering), c) understand the main topic words that the community is using by hovering over the dot and seeing the word cloud that shows up on the left. The core of our study lies in understanding semantic change and this visualization shows how this problem is associated with factors such as average age (which is coded in a set of brackets), gender composition. community size, closeness and Twitter's popularity metrics. For example, the user will notice that the larger the community size, the lower the lexical change, not only for all the categories but generally by each individual category. Same relationship holds for age brackets. These relationships can be seen in Fig. 1 in the first page of the paper.

4.4 Case study

One case study will help to understand the navigation and learning process in this visualization. Let us imagine a sports journalist that is covering martial arts this season and wants to understand this cluster of communities in order for his social media team to make an online strategy. In the first visualization they'll discover that the community "wrestler" is the largest in terms of followers and discovers popular twitter accounts such as @WWEUniverse and @WWERomanReigns.

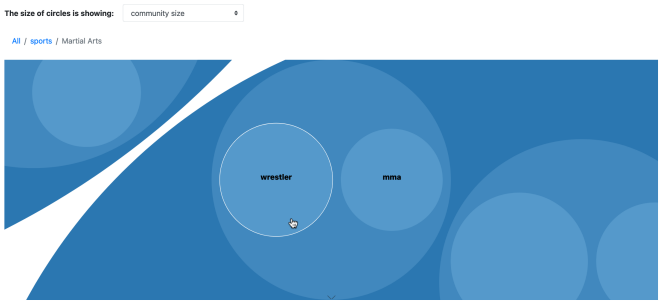

Fig. 2. Communities inside a bubble.

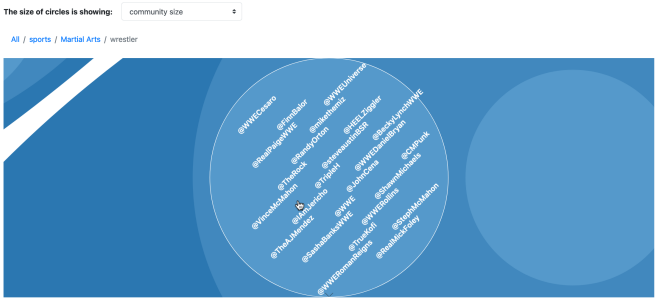

Fig. 3. Top followed Twitter account inside one community's bubble.

Next, they need to understand where the community is in terms of age so they can target their resources more effectively. They will learn in the second visualization that the average age of the community is located in the second quintile of the ranking, meaning it's relatively young audience.

Fig. 4. Users click a twitter account to explore in a new page.

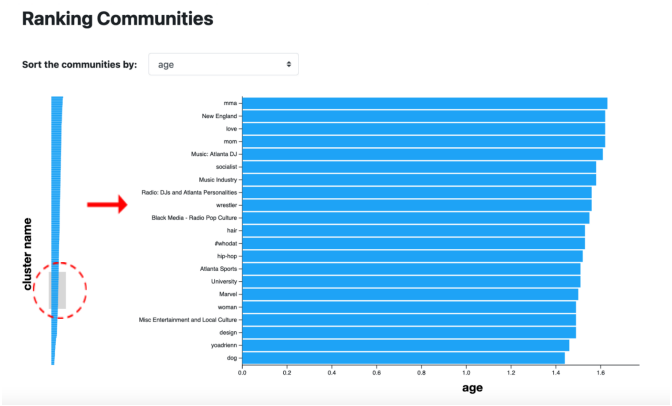

Fig. 5. **Community Ranking.** This tool allows the user to understand where the community is located in terms of a metric of preference

Finally, in the third visualization they will explore within the sports category where the wrestler community is located in terms of community size and semantic variation. In both dimensions, the community is situated on close to average levels of the total sports community. With this information in hand, the social media knows that they can trust that the main words of the community won't suffer many semantic variations and they don't have to be pending in real time on changes in the way people interpret those words. They'll discover that words such as #wwe, fan, xbox, game, etc are words that the community use, so that can help them do a text mining strategy to reach their target audience more easily. These results in our system show how this tool can help different specific actors of society to identify and understand the characteristics of online communities.

5 Discussion

5.1 Goal & Audience

Our main goal is that the user can understand the semantic changes in social media across Atlanta's communities and how the semantic changes are related to social factors. Examples of potential users are journalists, politicians, researchers, and society as a whole. For journalists and politicians, we hope to facilitate their policy making by allowing them to uncover and investigate potential social issues in the society. For researchers, we hope to provide them with this tool to explore the relationship between social factors, leading them towards a more rigorous analysis later. For the public, we hope this website is able to guide for current local trends online.

5.2 Novel Practices & Insights

New Practices There are three new practices that our visualizations provide. **First**, our circle packing visualization for community grouping (Figure 2 and 3) makes it possible for the user to explore the whole 149 communities interactively. Previous works [18, 22] use tables to describe each online community and only show a few selected examples, which is not comprehensive and difficult to read. The interactive design

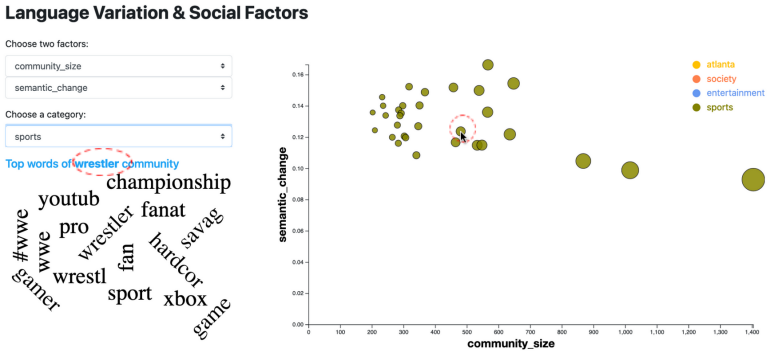

Fig. 6. **Inferring relationships.** With this scatter plot the user can understand the relationship between a pair of variables and can filter by category

in our paper allows the user to easily see the hierachy of communities and to click the top followed accounts in a community (Figure 4) to investigate what the community members love. **Second**, the horizontal bar chart with an adjustable range brush (Figure 5) allows users to choose a variable and rank the communities by that variable. Previous works [15, 22] have used static bar charts to show semantic changes across categories or groups. However, it is very difficult for them to use a static plot to show the ranking of 149 communities in a pdf page. Our interactive design not only resolves this issue with the adjustable range brush but also provides the variable dropdown menu to make it easy for users to compare communities with different dimensions (variables). For instance, ranking the communities by semantic change, age, gender will all yield quite different bar charts. **Third**, the interactive scatter plot enables users to explore the relationship between many pairs of social factors. The traditional visualizations [15, 18, 22] in the field typically use scatter plots to explore the relationship. However, they only provide a few plots due to the length limit of the paper and the readers are unable to explore the data. In this work, we allow users to choose any pairs of social factors and to filter communities by category, which can result in 1000+ combinations in total. Static visualizations are difficult to compare in that sense. At the same time, the wordcloud shown on the left when hover provides some descriptive information about each community. Our interactive design can significantly improve the user experience by integrating 3D data into the scatter plot (x-axis, y-axis, circle size) and by providing detailed community information through the tooltip.

New Insights We also have three levels of new insights from our visualizations. **First**, we explore individual communities with the circle packing visualization. For instance, we click into "random" community inside the "entertainment" category. With the help of this visualization, we can quickly identify the four communities there (streamer, local mix, porn, and games). To understand what is "local mix", I click it to the bubble and find people in this community follow accounts such as "cl_ atlanta", "ATL_Events". We then explore these two accounts and obtain a better understanding about what is "local mix" community. We did this to many communities, which helps us obtain a basic understanding of the Atlanta communities and correct some of annotations (community names) that is not accurate. **Second**, we use the ranking visualization to rank with different variables. For instance, we find out the "tik tok" and "BLM" communities have the lowest average age and "realtor" and "care" communities share the highest average age among their members. Similarly, we find "local celebrity" and "model" communities have the most semantic change from the norm and "entrepreneur", "trump", and "food" communities have the least semantic shift from the norm. **At last**, we use the scatter plot to identify factors that are visually correlated with semantic change. As shown in Figure 1, we can filter out bad variable and identify important factors (such as age, community size, closeness). Based on the exploration on the scatter plot, we further run regression analysis[4] to study the effect of social factors on linguistic variation in Atlanta communities in Table 4. We observe that community size, user favorite activity, and age have the most significant effects on the variation in all the models ($p < 0.001$). Among these three factors, age has the biggest impact, followed by community size and user favorite activity. Our results support the findings [8, 22] that **most linguistic changes happen in young ages** [26, 36]. We also provide evidence to the claim that **"small to medium" size communities tend to produce more lexical innovations than large communities**. Besides, user tweet activity and closeness show a lower level of significance ($p < 0.01$). So we also confirm that **users who are close to everyone in the Atlanta communities to have a closer language to the norm** [14]. Organization status is the least significant ($p < 0.05$). Two features (gender diversity and social status) are not significant in relation to linguistic variation in our study. Similar to previous findings [22, 38], we show that the category has no additional significant impact on distinctive language use beyond the user-based attributes. This suggest that *who* is involved in a community matters more than *what* they are discussing in a community [22].

[4]Details are in the Appendix C.

6 Future Work

The current overall visualization web application of this project works normally on desktops and laptops but lacks the support for mobile devices. On small mobile devices, the first two charts are functioning. Users can click to zoom in or zoom out in the circle packing graph and brush to zoom in or zoom out in the bar chart. However, hover event is not supported on some mobile device, so the word cloud is not showing and disappearing as designed in the scatter plot. The blocks are misplaced and not properly resized when this application is opened on small mobile devices such as cell phones. Although we use the CSS Flex layout module, the whole design, developing, and testing process is completed on desktops or laptops. In the future, we can modify the hover behavior and improve the overall layout to support small mobile devices, as most mainstream web and visualization designs.

Another future improvement is to integrate more hints for graphics. Now in the introducing part of this visualization, we have two paragraphs to explains some key variables in this visualization, and before each graphic, we have a short paragraph describing what is shown in the graphic and how to interact with it. Based on feedback, it is inconvenient when users want to reference some definitions in the introduction section. Because the paragraphs and graphics are different mediums and have unbalanced information density, some users may skip the paragraphs. A possible future work is to add step-by-step hints to lead users to click, hover or drag on the interactive graphics.

The size of the circle packing graph is responsive to the size of the web browser, but the arrangement of circles inside the graph is fixed. When the aspect ratio of the browser window is different from the ratio of the graph, we can improve it to prevent waste of spacing. The last possible future work is to add a module to detect if the Twitter accounts in the circle packing graph are valid. It not only prevents users from opening suspended accounts but also may provide interesting facts.

REFERENCES

[1] J. Androutsopoulos and E. Ziegler. Exploring language variation on the internet: Regional speech in a chat community. In *ICLaVE*, vol. 2, pp. 99–111, 2004.

[2] D. Bamman, C. Dyer, and N. A. Smith. Distributed representations of geographically situated language. In *ACL*, pp. 828–834, 2014.

[3] A. Blank. *Prinzipien des lexikalischen Bedeutungswandels am Beispiel der romanischen Sprachen*, vol. 285. Walter de Gruyter, 2012.

[4] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre. Fast unfolding of communities in large networks. *JSTAT*, 2008(10):P10008, 2008.

[5] M. Bostock, V. Ogievetsky, and J. Heer. D^3 data-driven documents. *IEEE transactions on visualization and computer graphics*, 17(12):2301–2309, 2011.

[6] C. Danescu-Niculescu-Mizil, R. West, D. Jurafsky, J. Leskovec, and C. Potts. No country for old members: User lifecycle and linguistic change in online communities. In *WWW*, pp. 307–318, 2013.

[7] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

[8] M. Del Tredici and R. Fernández. The road to success: Assessing the fate of linguistic innovations in online communities. *COLING*, 2018.

[9] M. Del Tredici and R. Fernández. Semantic variation in online communities of practice. *IWCS*, 2018.

[10] J. Eisenstein. What to do about bad language on the internet. In *NAACL*, pp. 359–369, 2013.

[11] J. Eisenstein, B. O'Connor, N. A. Smith, and E. P. Xing. Diffusion of lexical change in social media. *PloS one*, 9(11):e113114, 2014.

[12] M. Gavish and D. L. Donoho. The optimal hard threshold for singular values is 4/sqrt(3). *IEEE Transactions on Information Theory*, 60(8):5040–5053, 2014.

[13] M. Giulianelli, M. Del Tredici, and R. Fernández. Analysing lexical semantic change with contextualised word representations. *ACL*, 2020.

[14] A. E. Guinote and T. K. Vescio. *The social psychology of power.* Guilford Press, 2010.

[15] W. L. Hamilton, J. Leskovec, and D. Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. *ACL*, 2016.

[16] S. C. Herring and J. C. Paolillo. Gender and genre variation in weblogs. *Journal of Sociolinguistics*, 10(4):439–459, 2006.

[17] L. N. Hinton and K. E. Pollock. Regional variations in the phonological characteristics of african american vernacular english. *World Englishes*, 19(1):59–71, 2000.

[18] V. Kulkarni, R. Al-Rfou, B. Perozzi, and S. Skiena. Statistically significant detection of linguistic change. In *WWW*, pp. 625–635, 2015.

[19] W. Labov. *The social stratification of English in New York city*. Cambridge University Press, 2006.

[20] P. Li, B. Schloss, and D. J. Follmer. Speaking two "languages" in america: A semantic space analysis of how presidential candidates and their supporters represent abstract political concepts differently. *Behavior research methods*, 49(5):1668–1685, 2017.

[21] M. J. Lindstrom and D. M. Bates. Newton—raphson and em algorithms for linear mixed-effects models for repeated-measures data. *Journal of the American Statistical Association*, 83(404):1014–1022, 1988.

[22] L. Lucy and D. Bamman. Characterizing english variation across social media communities with bert. *TACL*, 2021.

[23] L. McInnes, J. Healy, and S. Astels. hdbscan: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11), mar 2017.

[24] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *Journal of Open Source Software*, 2018.

[25] D. Nguyen, A. S. Doğruöz, C. P. Rosé, and F. de Jong. Computational sociolinguistics: A survey. *Computational linguistics*, 42(3):537–593, 2016.

[26] D. Nguyen, R. Gravel, D. Trieschnigg, and T. Meder. " how old do you think i am?" a study of language and age in twitter. In *ICWSM*, vol. 7, 2013.

[27] D. Nguyen and C. Rose. Language use as a reflection of socialization in online communities. In *LSM*, pp. 76–85, 2011.

[28] D. Q. Nguyen, T. Vu, and A. T. Nguyen. Bertweet: A pre-trained language model for english tweets. *EMNLP: System Demonstrations*, 2020.

[29] H. Prichard. Linguistic variation and change in atlanta, georgia. *University of Pennsylvania Working Papers in Linguistics*, 16(2):17, 2010.

[30] G. K. Pullum. African american vernacular english is not standard english with mistakes. *The workings of language: From prescriptions to perspectives*, pp. 59–66, 1999.

[31] K. Rohe, T. Qin, and B. Yu. Co-clustering directed graphs to discover asymmetries and directional communities. *PNAS*, 113(45):12679–12684, 2016.

[32] D. R. Roth and A. Ambrose. *Metropolitan Frontiers: A Short History of Atlanta*. Longstreet Press, 1996.

[33] D. Schlechtweg, S. S. i. Walde, and S. Eckmann. Diachronic usage relatedness (durel): A framework for the annotation of lexical semantic change. *NAACL*, 2018.

[34] I. Stewart, S. Chancellor, M. De Choudhury, and J. Eisenstein. # anorexia,# anarexia,# anarexyia: Characterizing online community practices with orthographic variation. In *Big Data*, pp. 4353–4361. IEEE, 2017.

[35] I. Stewart and J. Eisenstein. Making" fetch" happen: The influence of social and linguistic context on nonstandard word growth and decline. *EMNLP*, 2017.

[36] S. E. Wagner. Age grading in sociolinguistic theory. *Language and Linguistics Compass*, 6(6):371–382, 2012.

[37] Y. Yang and J. Eisenstein. Overcoming language variation in sentiment analysis with social attention. *TACL*, 5:295–307, 2017.

[38] J. Zhang, C. Danescu-Niculescu-Mizil, C. Sauper, and S. J. Taylor. Characterizing online public discussions through patterns of participant interactions. *PACM HCI*, 2(CSCW):1–27, 2018.

[39] J. Zhang, W. Hamilton, C. Danescu-Niculescu-Mizil, D. Jurafsky, and J. Leskovec. Community identity and user engagement in a multi-community landscape. In *ICWSM*, vol. 11, 2017.

Corpus	TwitterATL
tweet count	7,602,853
token count	98,934,304
unique token count	750,033
# communities	149

Table 1. Statistics for TwitterATL datasets

A Data Collection

We took two approaches to identifying an Atlanta-based population: first, by querying the followers of a set of "seed accounts" (e.g., "navicenthealth", "civicatlanta") that we expected to have a large Atlanta-based followership, fetching the user objects for those followers, and filtering this set. Second, by scanning the user objects embedded in tweets on the "decahose", Twitter's 10% sample of tweets, and filtering this set. In both cases, the large set of user objects was filtered using a simple rule-based classifier on the "location" field to match against various conventional identifiers for "Atlanta, GA" (e.g., "Atlanta, Georgia", "ATL", "atlanta, ga", etc.)

This yielded 174,911 users we believed to be located in Atlanta. For each of these users, we then queried for the user's "friends" (i.e. the set of users that the target user follows.) On Twitter, who you choose to follow is the primary driver of the user experience, largely determining the tweets you see in your timeline. As such, the users you follow may reflect both your interests and social connections.

Note that both the follow graph and decahose based methods for assembling the Atlanta population have some bias – the follow graph method reflects the audiences of the seed set and thus is best served by a broad and diverse set of seed accounts, while the decahose method only reflects users who tweet (including retweets and replies.) Finally, the population only includes users who indicate their location.

A.1 Interest clusters

To study the linguistic variation of these Atlanta-based Twitter users, we first identify coherent interest clusters which reflect the myriad interests of this diverse population. Although there are many choices of who to follow on Twitter, users with similar interests may choose to follow the same or similar accounts. We use a decomposition method to uncover this latent structure and compute low dimensional vector embeddings for each user, such that users with similar interests will be near each other in the embedding space.

The identified Atlanta-based population consists of $174,911$ users who collectively follow $23,767,210$ distinct accounts. We first filter by considering only the subset of friends (i.e. the followed accounts) that have at least 1000 followers in this population, and then removing users (i.e. members of the population) who follow fewer than 10 of the remaining friends, yielding $147,049$ users and $5,603$ friends. We represent these follow relationships with a binary $n \times m$ adjacency matrix A_{ij}, with $A_{ij} = 1$ if user i follows user j, 0 otherwise. A is a large, sparse matrix, with $n = 147,049$ rows and $m = 5,603$ columns. We also rescale and regularize the adjacency matrix by the in- and out-degrees of the nodes, following the method in [31] (except we use the adjacency matrix rather than the Laplacian).

With this filtered and rescaled adjacency matrix A, we perform the truncated singular value decomposition (SVD) to obtain matrices U, S, and V^t, truncating to only the top k singular values of S to obtain U, an $n \times k$ matrix, with each row corresponding to a k-dimensional embedding of a user. We use the method of [12], which yields a truncation value of $k = 583$.

To obtain interest clusters, we next cluster the row vectors of U using the HDBScan Python package [23]. Rather than directly apply HDBscan to the $k = 583$ dimensional row vectors (or reduce k further via the truncated SVD if needed), we use UMAP [24] to reduce the embedding dimension following guidelines in the UMAP documentation.[5] Using a 10-dimensional embedding followed by HDBscan clustering, 60.1% of the population are clustered into 149 interest communities. Table 2 shows some examples of the identified communities, summarized by the top accounts followed by community members, and the top most distinctive words from the members' "bio text" fields. Although the interest communities are derived only from the follow graph, the coherence and interpretability of the bio text and top followed accounts supports our claim that we have identified meaningful interest structure for this population.

This method is perhaps more similar to Latent Semantic Analysis of document-term matrices [7] than spectral methods for network community detection, since the rows and columns effectively correspond to different sets of items. Community detection methods, such as the Louvain algorithm [4], typically focus on the link structure within a single set of nodes and may be better suited to revealing social rather than interest-based communities. In future work, we plan to investigate the linguistic features of such social communities.

A.2 Tweet history

Finally, using the Twitter API we assembled a dataset of roughly 12.5 million tweets (including replies and retweets) from the filtered, clustered Atlanta user population. This dataset consists of up to 100 recent tweets from each user in the population, which we fetched in mid February 2021.

B Semantic Change Detection Framework

To compute significant semantic shifts per community, we follow three steps: (1) generate token-level embeddings; (2) calculate semantic changes; (3) detect significant changes.

B.1 Generate Token-level Embeddings

Both datasets are from the Twitter domain. Therefore, we use the pre-trained BERTweet model [28] to extract linguistic features. This model is pre-trained on 845M tweets and has a better capability of representing tweets than the original BERT-base model.

In our study, we focus on two sets of glossaries: `DialectAAE` and `IdeologyLex`. `DialectAAE` has 80 AAVE-specific terms and is based on related works [17, 30]. Though not comprehensive, `DialectAAE` covers some reliable dialect variation in syntax, semantics, and orthography in AAVE. This word list enables us to validate and compare our method with other approaches on the TwitterAAE dataset. `IdeologyLex` is a list of ideological concepts and beliefs provided by [20][6]. This list allows us to study how people from different communities in Atlanta understand the same ideological concepts differently.

Apart from the two glossaries, we also compute embeddings for 1000 words randomly sampled from the most frequent 5000 words from each dataset. These words are used to construct a semantic change distribution for each community, which is later used to detect significant changes.

B.2 Calculate Semantic Shifts

For each target word, we calculate its average token embedding on one community by averaging its token-level embeddings across all the usages. For TwitterAAE, we simply compare AAVE and SAE communities by computing the shifts from one to the other. On the contrary, there are 149 communities for TwitterATL and pairwise comparisons would be too computationally expensive. Therefore, to measure how each community-specific dialect shifts from the norm, we randomly sampled K tweets from all the tweets to create the norm community, where K is the mean number of tweets across communities. Afterward, we compared all the communities against the norm community. The semantic shift for each term is computed with the cosine distance between two average token embeddings [13]:

$$\mathscr{D}(U_w^{c_1}, V_w^{c_2}) \quad = \quad 1 \quad - \quad \cos\left(\frac{\sum_{u_{w_i} \in U_{w_i}^{c_1}} u_{w_i}}{n_w^{c_1}}, \frac{\sum_{v_{w_j} \in V_{w_j}^{c_2}} v_{w_j}}{n_w^{c_2}}\right)$$

where $U_w^{c_1}$ indicates all usage embeddings for a term on the corpus c_1 and $n_w^{c_1}$ indicates the number of occurrences for the term w in the corpus

[5] http://umap-learn.readthedocs.io/en/latest/clustering.html

[6] After removing phrases, there are 133 ideological terms.

c_1. This score indicates the degree of semantic changes undergone by a word between corpus c_1 and c_2. As previous literature [33] suggests, this score does not distinguish a gain or loss of word meaning.

B.3 Robustness Check

Using the same method for sampling the control group norm community, we sampled 10 additional norm communities. Comparing the 149 interest communities with each of these norm communities and repeating the analysis yielded similar patterns of results.

B.4 Unsupervised Significant Change Detection

To detect significant changes, we construct a distribution of semantic changes with a set of top frequent terms. Specifically, we sample 1000 words from the most frequent 5000 words, which do not overlap with the two glossaries presented above. Based on this null distribution, we can estimate the probability of any target word's semantic change on a community corpus and determine if the change is significant or not through a right-tailed p-value test.

C Regression Analysis

C.1 Linear Regression

Based on the analysis above, we build a linear mixed (LM) regression model to study the effect of social factors on linguistic variation in Atlanta communities in Table 4. We observe that community size, user favorite activity, and age have the most significant effects on the variation in all the models ($p < 0.001$). Among these three factors, age has the biggest impact, followed by community size and user favorite activity. Besides, user tweet activity and network centrality show a lower level of significance ($p < 0.01$). Organization status is the least significant ($p < 0.05$). Two features (gender diversity and social status) are not significant in relation to linguistic variation in our study. Similar to previous findings [22, 38], we show that the category has no additional significant impact on distinctive language use beyond the user-based attributes. This suggest that *who* is involved in a community matters more than *what* they are discussing in a community [22].

C.2 Linear Mixed Model

We develop the mixed-effects model [21]:
`lexical_change` $\sim$ `community_size + age + network_centrality + user_fav_activity + (organization_status*social_status) + (1|category)`
on selected social factors. Specifically, we allow organization status and social status to interact with each other because personal and organizational Twitter accounts with many friends (high social status) are different users and should be distinguished from each other. Shown in Table 5, we demonstrate that **interaction between organization status and social status is significant** even though the social status alone is not significant and organization status is also not the most significant factor ($p < 0.01$).

Community Name	Comm. size	Top 10 (stemmed) Biography Words (1st row) / Followed Accounts (2nd row)
Emory Hospital (Health)	514	emori,health,@emorymedicin,research,medicin,@emoryunivers,fellow,care,diseas,professor,medic,md,resid,@winshipatemori,physician EmoryMedicine,emoryhealthcare,EmoryUniversity,emoryhealthsci,CarlosdelRio7,GradyHealth,CDCgov,ajc,EmoryDeptofMed,BarackObama
Gwinnett County (Atlanta)	355	gwinnett,mom,consult,realtor,duluth,clean,metro,pilat,profession,develop,help,atlanta,web,famili,counti GwinnettDaily,ajc,GwinnettNewsNow,wsbtv,GwinnettSchools,GwinnettChamber,GwinnettMag,GDPsports,GwinnettEvents,FOX5Atlanta
HBCU (Education)	417	morehous,colleg,spelman,student,educ,cau,@morehous,clark,communiti,univers,hbcu,offici,career,@cau,alumna Morehouse,SpelmanCollege,ajc,CAU,BarackObama,CityofAtlanta,HowardU,HBCUBuzz,wsbtv,CNN
Travel (Other)	511	travel,hotel,airport,trip,luxuri,experi,pilot,wanderlust,cruis,adventur,servic,compani,bed,airlin,food ajc,TravelLeisure,travelchannel,NatGeoTravel,Delta,CNTraveler,lonelyplanet,SouthwestAir,TravelMagazine,wsbtv
Democrat (Politics)	1883	#resist,polit,trump,liber,democrat,resist,vote,justic,#bidenharris2020,#fbr,#blm,retir,#theresist,wife,lover BarackObama,JoeBiden,KamalaHarris,maddow,ProjectLincoln,HillaryClinton,SpeakerPelosi,gtconway3d,AOC,JoyAnnReid
Christian religious (Religion)	1353	jesus,christ,husband,church,pastor,follow,father,love,christian,wife,god,grace,ministri,author,discipl AndyStanley,louiegiglio,JohnPiper,CSLewisDaily,ajc,lecrae,RickWarren,BethMooreLPM,plattdavid,christomlin
LGBT (Social Justice)	671	gay,lgbtq,lgbt,hiv,atlanta,communiti,activist,lesbian,queer,transgend,support,tran,aid,pride atlantapride,ProjectQAtlanta,theGAVoice,HRC,GAEquality,NOH8Campaign,glaad,BarackObama,TheEllenShow,HRCATL
Atlanta Falcons (Sports)	292	falcon,#riseup,sport,@atlantafalcon,fan,#falcon,atlanta,#brave,#truetoatlanta,nfl,brave,#hawk,#atlantafalcon,justwaldrop,#inbrotherhood AtlantaFalcons,M_Ryan02,debo,juliojones_11,devontafreeman,GradyJarrett,Keanu_Neal,VicBeasley3,TheFalcoholic,roddywhiteTV

Table 2. Examples of Atlanta online communities. Each has its subcategory in parentheses. On the right side, the 1st row contains the top 10 biography words and the 2nd row contains the top 10 followed accounts among members of a community.

Word	Community Name	Examples
defense	Atlanta Falcons	matt ryan can slice most **defenses** when he has better play calling on his side and almost every drive ends up in the red zone
	Norm	but in my **defense** you see me twice a year only at night time and only at a bar
class	Car	the e **class** initially stood for einspritzmotor german for fuel injection engine
	Norm	one of my teachers from animation mentor pete paquette was at blue sky when i was in his **class**
race	Science	shes talking about cultural differences not physical **race** differences
	Norm	i am from athens georgia for the **race** . we have participants from as far away as upstate new york
trust	Finance	if you have a **trust** did you actually fund it is your plan ready for the new secure act
	Norm	if you **trust** them i trust them
president	Democrat	when you took over **president** obama had done the hard work
	Norm	the **president** of stonehenge consulting group joined us today to share key success strategies
media	Trump	**media** to be so flagrant with their lies and misdirection
	Norm	open up a lucid **media** studio where ill be taking pictures and also booking gigs and helping other creators

Table 3. Ideological terms with examples used for different meanings in the selected and norm communities.

Variable	Estimate	Std
(intercept)	$1.012e-15$	0.048
community size	$-0.4529***$	0.050
social status	-0.0237	0.063
user tweet activity	$0.2983**$	0.091
user favorite activity	$-0.3721***$	0.097
gender diversity	0.0486	0.054
age	$-0.7046***$	0.100
organization status	$0.1871*$	0.073
network centrality	$-0.2339**$	0.075
category	0.0416	0.061
N	149	
R^2	0.685	
Adjusted R^2	0.665	

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 4. OLS regression results for the effect of community attributes on the fraction of words used in each community.

Variable	Estimate	Std
(intercept)	$-0.003***$	0.048
community size	$-0.442***$	0.042
user favorite activity	$-0.170***$	0.011
age	$-0.769***$	0.079
network centrality	$-0.178***$	0.047
organization status	$0.166**$	0.061
social status	0.012	0.012
organization status:social status	$-0.217***$	0.020
N	149	
R^2	0.668	

$^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table 5. LM regression results for the effect of community attributes on the fraction of words used in each community.