# A lexical approach to locating symbolic boundaries around cultural identities on social media

Zackary Okun Dunivin
Department of Communication, University of California, Davis
Department of Computer Science, University of California, Davis

August 23, 2024

**ABSTRACT**

Social groups and group affiliations are central to much social research even when research does not center identity. However, systematically identifying the symbols marking the boundaries around a particular group is often difficult, even for expert observers or group members themselves. I present an extension of a set of simple computational techniques to automatically extract large and highly specific lexicons (lists of words) representing a symbolic boundary around a social/cultural identity. The primary contributions of this approach are that it does not require high-quality labeled data and articulates challenges and affordances presented by social media trace data. Most importantly, the scale and scope of social media data often render ground-truth labeling of the identity of authors infeasible. To circumvent this problem, I use a small set of known identity signals to identify the social group of interest, which yields excellent results despite being a flawed and noisy representation of the group. Additionally, drawing on the computational grounded theory paradigm, I discuss the role of qualitative analysis in interpreting a symbolic boundary during and following the production of a lexicon. Finally, I address how computational methods can augment and expedite, rather than replace, traditional qualitative methods applied to text such as close reading and manual coding.

# 1 Intoduction

Boundaries are often invoked in the context of social science, particularly when dealing with cultural identity. This is not necessarily because boundaries are more likely to play a role in cultural processes, but rather that determining boundaries is often uncomplicated outside obviously cultural contexts. There are, however, a large set of situations where boundary delineation is a barrier to social research. In some cases, unspecified boundaries can be attributed to the nature of the data. More often than not, though, such ambiguity arises from properties of the system itself. The diversity, intersectionality, subjectivity, and fluidity of these boundaries imbue our social lives with richness and texture, and may also motivate social scientists as we endeavor to understand the processes governing the social. The same properties of social boundaries that ignite the sociological imagination (Mills, 1959) can impose enormous time investments if we are to derive scientifically adequate and meaningful accounts of those boundaries.

Historically, boundary specification in the social sciences has entailed qualitative methodologies such as participant observation, interviews, and content analysis (e.g. McCurdy and Uldam, 2014; Jorgensen, 1989; Watson and Weinberg, 1982; Kleinman et al., 1994; Gilbert, 2002; Lee and Roth, 2004). Symbolic boundaries are often implicitly located. The terms "role" and "identity" tend to be good indicators that research is centering boundaries rather than situations, behaviors, or dispositions. Over the past 30 years, workers in various disciplines have developed computational tools and methodologies to automate boundary specification. These approaches tend to fall into either of two camps: network analysis, which delineates the boundaries around groups of actors based on the relations that tie them to one another, and content or text analysis, which leverages trace data to elucidate the symbolic markers which impose or constitute those boundaries. As quantitative solutions to the boundary "specification" or "demarcation" problem were originally articulated in the context of social networks (Laumann et al., 1989), I briefly treat its history in networks before abstracting network methods as a specific case of classificatory solutions to boundary specification. I demonstrate why these methods are inadequate for a large set of cases of boundary specification which instead warrant a (computational) content analytical approach.

In this article I recontextualize and extend an existing method of content analysis, "frequency-based" lexicon extraction (Monroe et al., 2008), to handle especially difficult cases of boundary specification in textual records. Generally this can be understood as the problem of symbolic boundary specification in unlabeled data, i.e., data for which the social category or categories are unknown to the researcher. This is often the case for textual records generated by a large group of authors. Most often these data are sourced from social media platforms, but administrative records, parliamentary transcriptions, and journalistic or correspondence archives may also be "unlabeled" for the purposes of a particular research question. Methods for extracting group-specific lexicons yield the most reliable and high-resolution approaches to symbolic boundary detection in text. However, existing methods of lexicon detection demand labeled data, and are specified for cases where boundaries are easily separable, mutually exclusive, and limited to just several social contexts. Social media trace data by contrast are often very loosely structured with few formal boundaries. Social media data are further complicated by small (and thus low information) documents, which can make patterns difficult to detect, and viral text, which can erroneously amplify lexical patterns. The method I present here is designed specifically to meet the challenges posed by social media and other corpora comprising many murky boundaries. In a separate piece in this edition, a coauthor and I apply this method to locate symbolic boundary markers delineating the white nationalist discourse on Twitter (Dunivin and Lanigan, 2024). This case is subject to multiple, equally valid theoretical frames, among them social movements, dialogic/textual communities, and ideological/intellectual traditions. We encourage readers of this piece to refer to its companion for a full demonstration of the analytical process.

Finally, I situate this computational method of boundary specification in the social sciences more broadly. I follow Nelson's (2020) effort to rigorously articulate a sociological metamethodology of computational text analysis, which she calls "computational grounded theory" (CGT). By replicating grounded theory, an ethnographic method with a long history in the social sciences (Strauss and Corbin, 1997), CGT formalizes and emphasizes the role of human interpretation in otherwise quantitative analyses, unifying the reproducibility and precision of computational text analysis with the nuance and theoretic grounding of quantitatively derived concepts. In CGT, neither computation nor interpretation take primacy. Rather, they recursively inform one another throughout the process of developing the computational analysis. Here, I extend the jointly quantitative/qualitative ethos of computational grounded theory beyond the process of quantitative inference. Computational content analysis need not the beginning and end of a social scientific inference. Rather, content analysis should be regarded as independent of the quantitative/qualitative cleavage. Patterns extracted through computation can inform downstream analyses that follow either analytical paradigm. I am especially concerned with how qualitative practitioners understand the potential for computational tools to play a secondary role in social scientific inquiry. While Nelson characterizes deep reading as augmenting computational analysis, I suggest computation can augment qualitative analysis. I deliberately make this case as a response to valid concerns that computational social scientists (and funding agencies) have proposed computation as a replacement for qualitative analysis. Qualitative researchers can exploit the efficiency and breadth of computational processing

to facilitate their own qualitative analyses. By embracing computational methods, and perhaps even driving their development, practitioners of qualitative methods best position themselves to weather attacks on their paradigm and preempt charges of Luddism.

### 1.1 Symbolic boundaries and the boundary specification problem

Boundaries are an abstract concept referring to the properties of social and cultural categories or the existence of the categories themselves. Lamont and Molnár (2002) propound two types of boundaries, symbolic and social, which are useful in understanding different methods of boundary specification. They characterize social boundaries as "conceptual distinctions made by social actors to categorize objects, people, practices, and even time and space". Social boundaries are defined as "objectified forms of social differences manifested in unequal access to and unequal distribution of resources (material and nonmaterial) and social opportunities" or less critically "groupings of individuals" in the sense of social network clustering or modularity. Symbolic boundaries, through homophily and other mechanisms, instantiate social boundaries.

Knowing where these boundaries lie is a precondition to much social research. In some cases it is easy. Punks (Fox, 1987) and Hasidim (Tavory, 2010) are clearly identified by their manner of dress. University department or political party affiliation may be gleaned from organizational documentation or a simple survey. However, when social situations are complicated by multiple salient identities, poorly defined or unknown labels, and large symbolic vocabularies, locating social boundaries or symbolic boundary markers is nontrivial. Accordingly, multiple, diverse literatures have developed both explicit and implicit solutions to various boundary specification problems. Ethnography, interview, and close reading represent the bulk of the qualitative approaches to boundary specification. The quantitative side is represented by computational methodologies: network analysis, natural language processing, and machine learning. This review will treat only quantitative methods, though the literature on qualitative approaches is as great, if not greater.

Boundary specification is commonly treated as a classification problem. In social sciences, quantitative classification almost always relates to social boundaries: given a set of actors, which ones should be tagged with what labels? Classification may be supervised (we know the labels for a subset of the total population) or unsupervised (we don't know the labels, or aren't employing them to train a classifier.) We may seek a simple binary classification or look for multiple categories. If there are multiple categories are these categories overlapping or mutually exclusive? Lastly, the data used to impute classifications may be of qualitatively different types. Specifically, much actor classification in the social sciences is based on network data, a set of social relations linking actors to one another. Other approaches rely on non-network variables, most often demographic data, but also an actor's behavioral, interactional, or dispositional attributes.

The networks literature is a primary site of social boundary specification. Laumann et al. (1989) articulate a "boundary specification problem" in defining system boundaries in the context of social networks. In a network the size of the human population, where should we make the cut so as to get a subgraph that can offer a vantage on the particular facet of social life? I describe here a second order boundary specification problem. After defining a system, or having one defined for us by constraints in our data, can quantitative methods identify the symbolic or social boundaries intersecting the population? In the case of social boundaries, this is simply a matter of separating the group into distinct, typically nonoverlapping partitions. Locating symbolic boundaries by contrast entails finding the traits that define particular social categories.

Social boundary delineation occupies much of the network science literature. The network scientific approach is generally referred to as community detection, a diverse set of methods for partitioning network nodes into groupings based on network structure (relations between the nodes) (Danon et al., 2005; Fortunato, 2010). Community detection algorithms come in many flavors, each with strengths and weaknesses and particular structures which they are better suited to identify (Peel et al., 2017).

There exist also non-network methods for social boundary detection, generally referred to as classifiers. Clustering algorithms are unsupervised classifiers that specify a partition across the data without being given examples of particular classes. Common clustering algorithms include k-means or k-medians, random forest, DBSCAN, and Gaussian mixture models. Community detection can be thought of as an approach to clustering based on network structure. Supervised classifiers are typically some form of binary regression. In social boundary detection, the practitioner privileges the model's predictive capacity (the "computer scientific approach,") whereas in symbolic boundary detection, it is the strength of the association of each variable with the group which concerns the modeler (the "social scientific approach".) In other words the classification itself delineates a social boundary and the variables which are predictive of or correlated with a particular category can be thought to represent a symbolic boundary.

### 1.2 The frequency-based method of lexicon extraction for symbolic boundary identification

This study describes a method of identifying symbolic boundaries, or, more specifically, identifying a set of symbolic boundary markers from unlabeled data. This process produces a group-specific lexicon, a set of words that mark the symbolic boundary around a particular community. Monroe et al. (2008) present a thorough account of techniques and decision points for identifying differential word usage between two groups. The approach I present builds on a set of non-model-based approaches reviewed in Monroe et al. in two ways. First, their treatment assumes texts have been assigned high-quality labels denoting which group they belong to. The method I present here side-steps this problem by cheaply assigning low-quality labels to the text while still producing excellent results. Second, I address several challenges and affordances of social media and other big and situationally unconstrained data sets. By situationally unconstrained I mean that the data express a broad range of social situations and contexts. Political debate, organizational documents, and academic literatures are likely to be far more focused than the activity of a set of Twitter or Facebook users, because a single person's social media activity may cover all these in addition to remarks on their media consumption, relationships, mental health, etc. Further, social media is syntactically unconstrained, comprising formal and informal verbiage, including conversation, which leads to short documents and challenges in human and computer interpretation. I have no computational solution for the interpretive challenges, but do address some problems related to short documents.

Monroe et al. argue for model-based approaches for lexicon extraction to the exclusion of non-model-based approaches. They assert that their Bayesian models provide point estimates for odds ratios. However, their treatment is specified for cases that look fairly different from the social media data considered here. In part this is because U.S. Congressional debate corpus they analyze contains less group-signalling jargon, which their results repeatedly demonstrate. But the more salient difference between Monroe et al.s' formulation and the Twitter data I used to develop this approach is the enormity of trace data ($10^8$ records, $10^6$ authors.) The combination of these two factors means there are many more robust signals in these data than in Monroe et al.'s. Therefore we can use much more stringent criteria to admit a symbol to our lexicon, and still end up with large and thorough, though not necessarily complete, set of boundary markers.

The Bayesian approach is ill suited to these data for several reasons. The most serious is that without high-quality labeled data, point estimates for differential term usage are not meaningful in a strictly statistical sense. An alternative approach could be to manually apply in- and outgroup labels. But this raises not only the efficiency bottleneck, but also difficulty in human interpretation of short documents and the fact that even a committed ingroup member may dedicate only a small fraction of their online behavior to group-related discourse. This is further complicated because commitment changes over time, which means that while ruling in might be easy, type II error will be high if based only on a small sample of an author's total output in the corpus.

## 2 Extending the frequency-based method for large, unlabled data

This study expands upon a set of computational methods for identifying a large set of group-identity symbols in textual records. As reviewed in Monroe et al. (2008), the method advanced here is a "frequency-based" approach to lexicon extraction. The frequency-based approach employs odds ratios to evaluate differences in term use between two groups, but does not rely on regression modeling to estimate statistical significance.

I present two developments to frequency-based methods to account for some challenges presented by social media trace data. First, I explain the importance of removing multiple copies of the same document, which account for a very large fraction of documents on platforms such as Twitter and Tumblr. More importantly, I describe a process for producing high-quality group-specific lexicons in cases where the group affiliation of utterances or authors is unknown. Table 1 gives a overview of this process.

### 2.1 Preprocessing the text: Removing copied utterances and identifying phrases in the corpus

Preprocessing text is an important step in any quantitative text analysis. Exactly which transformations are warranted is determined largely by the analytical algorithm(s) we will run over the text and to a lesser degree by particularities of the text itself. Common preprocessing operations include converting to lower case, removing punctuation and stop words, partitioning text into sub-units (such as paragraphs or sentences), and stemming words. Setting characters to lower case and removing punctuation is necessary to identify terms that differ in use only due position in a sentence. Similarly, stemming can be useful for amplifying particular signals and organizing the lexicon.

Several additional preprocessing steps are critical before applying the method. The first, n-gram extraction, is common to many tasks and data. N-gram extraction refers to the identification of word sequences of length $n$, typically bigrams and trigrams, that occur at much higher frequency than would be predicted from the frequency of the individual components. Since lexical identity symbols can be quite specific, it is imperative that we identify common phrases in

Table 1: Overview of the general methodological approach

| Step | Description |
|---|---|
| 1 | Select data that contains high density of a given ingroup and relevant outgroups |
| 2 | Select words used only by the ingroup |
| | • Small set of *very specific* terms. |
| 3 | Identify all authors who know these words (or use them frequently) |
| 4 | (Optional) Repeat Steps 2 and 3 for a particular outgroup |
| | • Otherwise, outgroup is the set of all authors outside the ingroup. |
| 5 | Find the words that are used by these authors and not the outgroup |
| | • Rank by odds ratio |
| 6 | Manually verify the list of words extracted in Step 5. |
| | • If low-quality reiterate from prior step. |
| | • If high-quality, perform inquiry incorporating the lexicon. |

our data. Alternatively, we may include all bigrams and trigrams in our set of words for a greater computational cost. The second additional preprocessing step is peculiar to social media, Twitter (among others) in particular. This is the removal of viral text. Platforms with sharing features (e.g., Twitter, Facebook, and Tumblr) are abundant with copied text. A document which has been copied many times in a data set is likely to unduly amplify particular rare words that happen to occur in that post. Since it is likely that members of the same group are closely linked in a network and are (multiply) motivated to share similar content, this has the effect of boosting the frequency of the words in a viral post within the ingroup and likely not in the outgroup. We must therefore screen out copied text prior to analysis. In fact viral amplification can similarly bias the results of n-gram extraction, thus copied posts should be excluded from the set of posts used to identify phrases.

Importantly, much of this preprocessing should occur prior to locating the in- and outgroup, since labeling is based on the use of particular words. Preprocessing is primarily intended to merge and amplify signals that would otherwise be distinct due only to differences in morphology, i.e., words that differ only in number, tense, part of speech, etc.

### 2.2 Locating the ingroup in the data

This method was developed specifically for unlabeled data. There are several potential barriers to labeling data. The primary impediment is that curating training data is expensive. The time it takes to label data is a huge burden, which increases with the difficulty of classification problem, i.e., the strength and frequency of the predictive features. In the case of labeling for lexicon extraction, tagging individual documents require an enormous data set, because of the number of features (words) and each word feature's relative infrequency in the corpus. Labeling training data must be performed by experts who can accurately identify the classes of interest. For many problems, such as image recognition, average adults are suitable experts. This is the intuition behind most crowdsourcing, especially through Amazon's Mechanical Turk. In the case of social identity labeling, many groups require specialized knowledge that precludes crowdsourcing. Moreover, multiple, dynamic identities complicate the process of tagging individual authors, increasing the likelihood of false negatives. Perhaps the actor only embodies a particular identity for a short period of their total period of activity, or draws on that identity consistently throughout their history, but only in a small fraction of their total activity, which may include a large absolute number of posts, indicating a robust, if less salient identity.

An additional problem in classification is that an identity that is active in a particular post may be difficult to detect. Social media data presents two sources of detection difficulty. The first is that documents, e.g., tweets, tend to be very short. Limited context makes deducing the exact meaning of a post difficult. The second is that the Internet is rife with irony (Merrin, 2019; DeCook, 2020). Even with a fair amount of context, pinning down the exact meaning(s) of 140 (or even 280) characters can be extremely tricky. In practice, these challenges entail throwing out a lot of data that may be informative, which means that not all of the labeling hours actually translate to training data hours. A further source of wasted effort is the large proportion of posts that may be completely irrelevant to demarcating the in- and outgroups we wish to distinguish. Posts about movies, sports, or a day at work may not signal anything pertinent to the identity of interest.

One solution to the challenges of labeling such diverse data as social media records is to pursue an unsupervised approach, which does not rely on labeled data. Unsupervised learning detects correlations between features, and partitions entities based on highly correlated values across subsets of features. In this case the features we train on and the entities we wish to distinguish are identical. We care about the sets of correlated feature-values, here (crudely) the presence or absence of a word, rather than the partition over the set of documents or users. Perhaps the most obvious way to partition users or authors is to identify the group from whom we intend to extract a lexicon, and then examine which features (words) the algorithm used to draw the boundary. A more sophisticated approach might bypass the categorization of documents and authors entirely to find associations between sets of words. In text analysis, this is commonly accomplished by topic modeling. Topic modeling achieves mixed results, often struggling with fine-grained distinctions, and is especially difficult on small documents, such as tweets. An alternative unsupervised approach might be to apply a clustering on word embeddings, which are themselves unsupervised features attached to words based on co-occurrence. While such an approach may in fact identify one or multiple partitions of words representing the group of interest, results are far from guaranteed.

The preceding paragraphs outline the challenge posed by labeling training data and uncertain results achieved by unsupervised methods. The approach presented here circumvents both problems by automating the process of applying low-quality labels to the data set. In reality, it circumvents only the problem of the expense of human coding, sacrificing coding accuracy. However, as I will explain later, the following steps of the method are such that we can be largely unconcerned by this rather "lazy" solution.

The remainder of this section describes an automated process for extracting low-quality labels in textual data. This process is analogous to a single-step snowball sample to identify a community from network data. Rather than identifying authors as seeds, we select a set of "seed" terms, which are highly correlated with the group from whom we seek to extract a lexicon. Having selected these terms, we can identify all the authors who have used these words, effectively taking one step on a bipartite network of authors and terms. More stringent requirements may be set, such as knowing all terms, or setting a threshold for "knowing" that is greater than 1, e.g., using the term 10 times. In subsequent steps in extracting the group-specific lexicon, we might opt to exclude authors who meet the simple criterion (knowing one word), but not more stringent ones, as neither reliably in the ingroup or outgroup.

The set of seed terms need not be large, between 3 and 10 should be sufficient, but the chosen terms should occur fairly frequently in data. There is no hard-and-fast rule about how frequent is frequent enough. The practitioner must exercise discretion in determining a sufficient set of words to capture most of the ingroup in the sample while screening out relevant outgroups. Having a good sense of proportion ingroup members in the sample can facilitate this judgment. Similarly, having selected seeds terms, determining the parameters for inclusion in the ingroup or outgroup based on the strength of association between author and seed terms requires informed intuitions about the relation between identity and term use. Ultimately the quality of the resultant lexicon will indicate whether the seeds and parameters were appropriate.

The approach to locating the ingroup within our data is equally suited to locating an outgroup. If we opt not to select an outgroup, the outgroup defaults to every author in our sample who has not been identified as a member of the ingroup or removed from the sample due to ambiguous membership in the ingroup. It is important to consider who constitutes the outgroup, as they should be as similar to the ingroup as possible apart from those terms that truly demarcate the ingroup from the outgroup. Identities are multiple and interdependent. In a simple sense of interdependence, covariance, I mean that the presence or degree of one identity in an actor informs the probability of each other identity.For example, a social identity that is dominated by men, for instance hunters or woodworking enthusiasts, will have higher rates of terms relating to sports than the general public, because men are much more likely than women to discuss sports. Covariance between multiple identities cannot be completely controlled for. However, we can do our best to mitigate its effects by choosing an outgroup that differs from as the ingroup as little as possible, while still maintaining a hard distinction with respect to the identity we are trying to characterize. Qualitative investigation of the lexicon helps to identify covarying identities and subsequently manually remove words or categories of words representing those identities from our lexicon.

### 2.3 Extracting a lexicon by ranked odds ratio

Having defined the in- and outgroup of authors and preprocessed documents, we can proceed with making the relative frequency calculations. A similar account of these calculations is given by Monroe et al. (2008). As with most steps in this process, the practitioner must apply their own judgment at multiple decision points, based both on intuition and an iterative process of lexicon extraction and parameter refinement, as described by Nelson (2020.) However, the results are frequently robust to a fairly broad set of specifications, as multiple parameter settings can produce very similar lexicons.

This method identifies a set of words that are associated with a particular group identity. Thus it compares some sort of strength of association (here a frequency) of each term with some group of interest against some relevant outgroup. Term frequencies are calculated by counting of the instances of $N$ terms for each of the two groups, $\tau^{in}$ & $\tau^{out}$. In its simplest formulation, this would be raw count of each term. However we might want to count the number of authors who know (i.e., have used) each term, or the number of authors who have used a term at least some number of times. I have found that number of authors who know a term yields the best results, as it prevents a small set of authors from inflating the counts; taking a logarithm of the each author's term counts would likely have a similar effect. Once we have counts for each term, we need to scale the counts to account for population size and productivity. The simplest way to do this is to divide the counts by total number of words in each group, which simultaneously controls for group size, number of documents, and document length. After scaling the counts we divide the ingroup frequency by the outgroup frequency to calculate the relative frequency. Equation 1 describes the calculation of the relative frequency vector, $\varphi \in \mathbb{Q}^N$. A complementary algorithmic account is given by Algorithm 1. This equation applies to any of the previously described methods of calculating the term count vectors, $\tau \in \mathbb{Z}^N$, and sum of author productivity vectors, $\omega = \sum_{i=1}^{k} \pi_i \,|\, \pi \in \mathbb{Z}^k$.

$$\varphi = \left( \frac{\tau_i^{in} \omega^{out}}{\tau_i^{out} \omega^{in}} \right)_{i=1}^{N} \tag{1}$$

Following the process outlined above is likely to identify very high relative frequencies for rare words that are not salient boundary markers due to their infrequency. In order to screen out infrequent words from a lexicon, we can impose a threshold such as total number of authors who must know the word in order for it to be included in the lexicon.

**Algorithm 1:** ODDS RATIO For a set of terms, calculates the odds of the ingroup's use of each term relative to some outgroup's

**Input:** A set of terms, $T = \{t_1, t_2, \ldots, t_N\}$, a pair of sets of authors, $A^{in}$ and $A^{out} \mid A = \{a_1, a_2, \ldots, a_k\}$, and pair of 2D matrices, $C^{in}$ and $C^{out} \mid C = \begin{bmatrix} c_{11} & \cdots & c_{1N} \\ \vdots & \ddots & \\ c_{k1} & & c_{kN} \end{bmatrix}$, where $c_{ij}$ is the number times the author, $A_i$, has used the term, $T_j$.

**Output:** A vector, $\varphi$, of the odds ratio of use of each term for all authors in $A^{in}$ relative to all authors in $A^{out}$ scaled for the productivity, $\omega$ , of all authors in the respective groups.

$\varphi \leftarrow T * 0$ (initialize the relative frequency vector)

$\omega^{in} \leftarrow \sum_{j=1}^{k^{in}} C_{,j}^{in}$ (total number of words uttered by the ingroup)

$\omega^{out} \leftarrow \sum_{j=1}^{k^{out}} C_{,j}^{out}$ (total number of words uttered by the outgroup)

**for** $j \leftarrow 1$ **to** $N$ **do**

- $\tau^{in} \leftarrow \sum_{i=1}^{k^{in}} C_{ij}^{in}$ (count of ingroup term use for all terms)
- $\tau^{out} \leftarrow \sum_{i=1}^{k^{out}} C_{ij}^{out}$ (count of outgroup term use for all terms)
- $\varphi_j \leftarrow \frac{\tau^{in} \omega^{out}}{\tau^{out} \omega^{in}}$ (scale by author productivity and calculate odds ratio for each term)

**return** $\varphi$

### 2.4 Qualitatively validating a candidate lexicon

The description of this method has thus far detailed the acquisition and preprocessing of data and the calculation of relative frequencies for every word that occurs in the data set. These relative frequencies are intended to indicate how much more often a particular group uses a word compared to some outgroup; a subset of the words are to be assumed to represent a boundary separating one particular group from either a particular or general other. The efficacy of the configuration of a particular data processing pipeline, i.e., the choices the researcher makes at each of the many decision points detailed above, must not be taken for granted. Instead the researcher must evaluate extracted lexicons (and their parameterizations) through their qualitative understanding of the social context.

The set of words and corresponding relative frequencies does not itself constitute a potential lexical boundary. The words must be ranked by relative frequency and those with the greatest differential expression taken as a boundary around a social category in question. The first and most important of these steps is to rank the words by relative frequency and

read through them in descending order. The words at the top of this list should represent the strongest group identity signals (by definition the seed terms, as they completely separate the two groups), followed by progressively weaker ones. The opposite pole represents terms that separate the outgroup from the ingroup, though, depending on how the outgroup is defined and sampled this may not be a substantively informative boundary. The most import thing here is to ensure that the list represents the boundary. The top of the list should contain only terms that crisply demarcate the ingroup, but as we move further down, good signals may become intermingled with ambiguous ones or indicators of other identities due to multiple identity covariance that was not controlled for in the outgroup definition. It is likely too that the list will be somewhat stratified by topic, with overlapping bands that represent various facets of the identity. If this pattern does not emerge, and instead the terms seem too general, represent a different boundary, or the relative frequency differences seem too small, this is an indication that at least one step was poorly specified. It may be that the in- or outgroup were not defined well, the terms in the relative frequency calculation need to be tweaked, or even that the data do not actually capture the groups they were intended to. It is easiest to try tweaking the relative frequency calculation, since it is the furthest downstream in the process, and the hardest to intuit.

The importance of qualitative interpretations of the social phenomena in question cannot be understated. The complete failure of the frequency calculations to extract a group-specific lexicon is often so easy to identify that invocation of qualitative observation or theory may go completely unnoticed by the researcher even as they make it. In cases where the candidate lexicon is generally more coherent, the qualitative interpretation becomes more explicit, as the evaluation of the lexicon is more difficult, and thus demands deliberate and nuanced hermeneutics. Aberrant words may be erroneous, indicating that the extraction process was poorly configured. Yet surprising phenomena are often precisely what we hope to flush out through inquiry. Distinguishing fallacious associations from unexpected discoveries is unavoidably fraught, and is often insoluble through quantitative inference. Close reading is the best tool for resolving such a dilemma. We can, however, let the computer assist us with this step through targeted close reading. In a targeted reading, the researcher mines instances of the term or terms in question in order to quickly access many utterances containing the term. This targeted reading may lead to further investigation of particular authors or other terms that can provide more context for the role of the original term or some discursive category it may be taken to represent. Such investigations serve a dual purpose. In addition to the immediate goal of identifying a lexical boundary, the researcher also deepens their understanding of the social phenomena in question, as in a non-computer-assisted close reading.

Qualitative investigation may solidify the researcher's confidence in a particular set of terms, but more often than not it raises questions or concerns that inspire another round of parameter selection and frequency calculation. Progressively the researcher hones in a suitable lexicon, while also deppening their qualitative analysis of the subject. Nelson's (2020) computational grounded theory elegantly describes this iterative process for "efficient, rigorous, and fully reproduceable" text analysis. The preceding paragraphs correspond to the "pattern refinement" step of her method (see Figure 1, in which deep reading is employed to verify the success of the lexicon extraction.

## 3 Discussion

This paper extends the frequency-based toolkit for group-specific lexicon extraction. Prior formulations of the approach require that authors or documents be correctly labeled as belonging to the groups of interest. In the case of very large data sets, labeling by hand is usually infeasible and even when manageable may offer only a small subset of the data, eschewing the benefits of big data. My primary methodological contribution lies in arguing that low-quality automatic group labels can be substituted for high-quality manually coded ones, which is demonstrated by the accompanying empirical study (Dunivin and Lanigan, 2024). The results presented there as well as the full lexicon given in the appendix show that large and comprehensive lexicons may be generated through the process I describe here. Bolstering support for computational lexicon extraction, this list is much more comprehensive than any that could be generated through close reading or crowd-sourcing via survey. Compared with earlier applications of the frequency method on relatively small corpora, the big data approach obviates the need for a statistical test for significance as advocated by Monroe et al. (2008), as very large samples with a simple frequency threshold prevents erroneous associations. Even with Bonferroni correction, the p-values in the accompanying lexicon are astronomically small and still meaningless.

Readers may wish for a more rigorous demonstration of the method than the paired empirical study on White nationalists. Ideally, this would be shown by robustness tests whereby lexicons derived using ingroup's defined by different sets of seed terms are compared to results of ground-truth labels. Nevertheless, the scale and diversity of these data, which are the greatest virtue of social media data, prohibit ground-truth valuations. More to the point, the grounded theory approach builds in robustness by leveraging deep knowledge of the social identity in question developed by the researcher through means other than computational analysis to validate the quality of a lexicon. While the final results of analyzing the lexicon should yield novel understandings to the research community, very little should be surprising to those conducting the research; their understanding of the phenomena involved should have largely been developed through close reading conducted prior and during lexicon extraction. The primary value of this automated approach

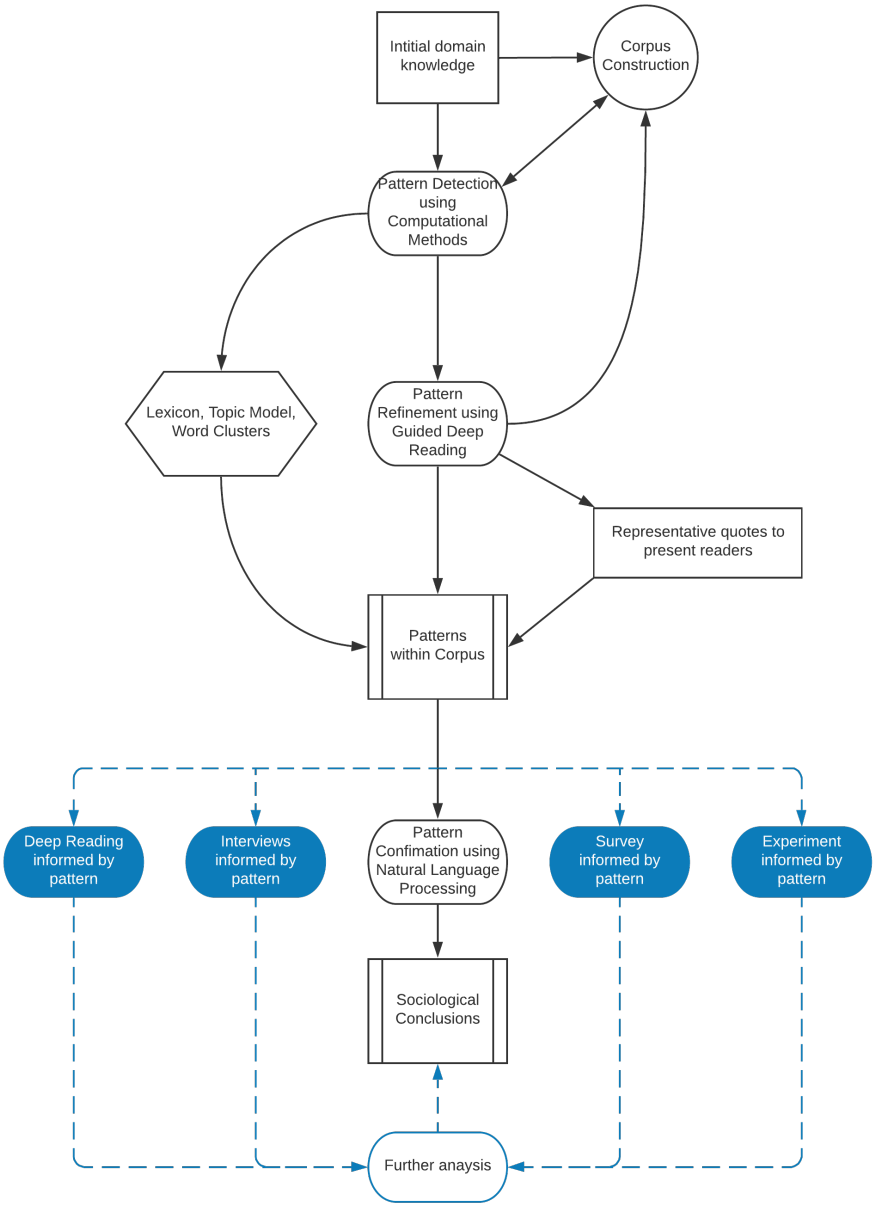


Figure 1: Extension of Nelson's (2020) computational grounded theory framework. Shaded and dashed structures (blue) represent proposed alternative processes beyond the CGT framework. All other structures (black) represent a slightly modified version of the original figure.

to symbolic boundary detection is that it yields more expansive results than manual approaches to lexicon extraction. The lexicons this method generates greatly exceed the number of terms compared those which could be derived by close reading alone. They may however include some additional thematic categories, as these themes may not have been apparent to the researcher through close reading, and were only visible through assemblage into a lexicon. In the study of white nationalists, I discovered a set of "lexical blackflace" terms. Many of these terms were known to me, but I would not have identified them as an important component of the white nationalist discourse without seeing them in aggregate. Moreover, considering these terms independent of their context made clearer that they were not simply anti-black racism, but specifically invoked to mock and dehumanize speakers of African American Vernacular English (AAVE). This illustrates how close reading the computational output of the lexicon yielded a more thorough account of white nationalist discourse than did close reading the tweets alone.

### 3.1 Computational qualitative analysis in and beyond boundary specification

I characterize this method as an automated method of lexicon extraction. However, the automatic component is actually just a small part of an iterative process wherein the results of the lexicon extraction must be validated within the researcher's intuitive understanding of the social phenomena in question. Nelson's (2020) computational grounded theory articulates this process in detail (summarized in Figure 1). In computational grounded theory, the parameters used to define the lexicon are refined until substantial portions of the lexicon are brought in line with intuition and erroneous markers that can be explained away are no longer present. Nelson stresses that computational grounded theory is more rigorous and reproducible than traditional content analysis. Part of the "rigor" or "objectivity" of the approach is that the computational tools do not enable piecemeal excision of aberrations. Since the parameters are specified at a very abstract and theoretically removed level, targeting very particular challenges to theory for excision is difficult. If for instance an iteration of the lexicon extraction produces a subset of words that are theoretically problematic, say knitting hobbyists appear to be frequently discussing Cajun food, we must either justify the presence Cajun cooking in our account of the identity of knitters, or make some adjustment to the extraction process. This can be done either by tweaking the parameters for lexicon extraction or redefining the set of actors we take as the in- and outgroup. The parameters of extraction used here (and those of any other computational text processing method) are very distant from concepts such as Cajun cooking, so any solution that removes Cajun food must be very general, resulting in widespread changes across the lexicon.

This article begins by framing group-specific lexicons as a symbolic boundary around a social category. The framework of boundary-specification problems, while originally articulated in the context of networks and social boundaries (Laumann et al., 1989), provides a scaffolding for understanding the problem of specifying symbolic boundaries as the flow of decisions around system representation (data acquisition), system description (lexicon extraction), and system analysis (inference using the lexicon) highlight how boundary specification is in fact a central problem in social research, and one that makes the same demand of the researcher across problems. "Where should I draw the boundary around this system so as to capture and make tractable the social phenomena I am trying to understand?" This presents several interrelated problems:

- Location: Was the boundary drawn around the relevant set of actors?
- Scope: Were the relevant details recorded given a particular set of methods?
- Scale: Is the system representation (the data) large enough to ensure conclusions are representative of genuine patterns?

Boundary specification, both here and in its original context, is generally a problem of the first type. However, it should be plain that these each of these questions must be resolved in order to execute any social inquiry. Happily, the answers are often obvious. But in many cases (more often than most believe) we encounter non-trivial, even prohibitively recalcitrant problems. Social research necessitates observational and theoretical intuitions around social situations. Science, whether conceived through a positivist or constructivist frame, entails recursive processes of observation and theorizing. This recursion is typically understood as a collective process of knowledge production, but it can be found throughout the nested hierarchy comprising rigorous inquiry. Boundary specification represents a fairly low-level recursion, wherein the boundary is validated by intuition while at the same time challenging it.

Nelson succeeds in articulating this recursive relationship, demonstrating that contextual understandings based on theory and qualitative analysis are essential to responsibly undertaking a (nominally) computational analysis. However, she characterizes computational text processing as ultimately building toward quantitative analysis. This portrayal of computational tools as quantitative in nature greatly undersells the utility of computation to facilitate textual analyses. In particular, lexicons such as the one derived here can guide the development of a broad variety of quantitative and qualitative studies other than verifying patterns extracted through computation. These lexicons can be used to inform interview questions or strategies, facilitate participant observation, construct a survey, design an experiment, or

contextualize conversation analysis or other deep readings. Figure 1 situates the process of computational grounded theory in a much broader methodological field.

On its face, the characterization of computational text processing as a quantitative method is sensible. Computers are fundamentally quantitative entities, adding and subtract bits of information. We teach computation as a mathematical or engineering discipline, and, as social scientists, conveniently calculate regression analyses with them. Nevertheless, most of what we do with computers is not "quantitative": we write emails and blogs, we read the news and books, we do graphic design and video editing, we record and generate music. As social scientists, we use computers to conduct experiments, give surveys, code text and interviews, and scrape trace data, none of which we characterize as quantitative.

If so much we do with computers is not quantitative, why do we treat NLP as a quantitative paradigm? A category error emerges when we recognize that there is an implicit understanding that quantitative methods are inherently inferential or necessarily feed into a quantitative inference. Inference is, however, only one of a number of steps in scientific inquiry. Data must be acquired and processed, and then described or analyzed.While many computational tools deliver analysis, most are merely processors. Accordingly, we refer to natural language *processing* rather than *analysis*. Techniques such as topic modeling, vector embeddings, knowledge graphs, and the approach advanced here are not in themselves inferential tools. Rather they are (quantitative) approaches to categorizing or organizing words. Once processed, we may feed this organization into a quantitative analysis à la Nelson, or qualitative one such as the paired study on white nationalists demonstrating the application of the method. Alternatively, the organization may not lead immediately to analysis at all, and rather be employed to collect or process new data. Quantitative analysis need not be the ultimate destination of the inquiry.

### 3.2 The future of computational qualitative analysis in a world with LLMs

The space of computational qualitative analysis has dramatically expanded with the advent of large language models (LLMs) (Brown et al., 2020). The natural language output of LLMs and unprecedented interpretive and reasoning capacities enabling zero-shot learning offer many new possibilities for qualitative research aided by computational tools. The most immediate of these is the application of human-generated qualitative codes by an LLM to classify large samples of text (Chew et al., 2023; Dunivin, 2024). While this is an exciting new application of machine learning that has eluded previous supervised learning attempts, they cannot replace close reading by a trained human researcher to determine the relevant qualitative categories. The frequency-based method of lexicon extraction can aid in this process and cannot be replicated by an LLM for several reasons. Critically, the LLM likely lacks sufficiently deep knowledge of identity in question, especially if it represents a fringe group. Moreover, the patterns of group identity are typically so sparse in the data that even with an arbitrarily large context window, an LLM would fail to identify many or any of them. Finally, an LLM, while considerably cheaper than an Mturker or undergraduate student, is still much too expensive to process millions of documents. I am greatly enthusiastic about the prospects of LLMs for new avenues of qualitative research in identity and beyond, yet their limitations ensure many simple NLP methods such as this process for lexicon extraction will remain useful for the foreseeable future.

## 4 Conclusion

Personal computing, improvements in computational processing power and memory capacity, the Internet, archival and bureaucratic digitization, and social media have coalesced to provide both an unprecedented volume of social behavior records and the means to process those data. These technological and social-structural transitions far outpace the impact computational methods have had on content analysis, or academic knowledge production more generally. Computational text processing and analytic methods such as the one advanced here have the potential to enable inquiry which is heretofore intractable to existing methods and facilitate traditional content analysis by improving breadth and efficiency. This is particularly true in the case of social media data, which give a window into the daily activity people use to construct their identities and give meaning to their lives.

Natural language processing presents a tremendous opportunity to integrate human interpretative and synthetic abilities with mechanical speed, reliability, and precision. However, reconsideration of existing quantitative paradigms will likely reveal that theorizing and qualitative observation are in fact operating throughout so-called quantitative processes. Recognizing and explicating the mutually informative relationship of quantitative and qualitative approaches can only improve the rigor of our science, and, one can hope, continue to break down barriers between paradigms, leading to more innovative and impactful research.

## References


Brown, Tom, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems 33*, 1877–1901.

Chew, Robert, John Bollenbacher, Michael Wenger, Jessica Speer, and Annice Kim 2023. LLM-assisted content analysis: Using large language models to support deductive coding. *arXiv preprint arXiv:2306.14924*.

Danon, Leon, Albert Díaz-Guilera, Jordi Duch, and Alex Arenas 2005. Comparing community structure identification. *Journal of Statistical Mechanics: Theory and Experiment 2005*(09), 9008–9017.

DeCook, Julia R. 2020. Trust Me, I'm Trolling: Irony and the Alt-Right's Political Aesthetic. *M/C Journal 23*(3).

Dunivin, Zackary Okun 2024. Scalable qualitative coding with LLMs: Chain-of-thought reasoning matches human performance in some hermeneutic tasks. *arXiv preprint arXiv:2401.15170*.

Fortunato, Santo 2010. Community detection in graphs. *Physics Reports 486*(3), 75–174.

Fox, Kathryn J. 1987. Real Punks and Pretenders: The Social Organization of a Counterculture. *Journal of Contemporary Ethnography 16*(3), 344–370.

Gilbert, Linda S. 2002. Going the distance: "Closeness" in qualitative data analysis software. *International Journal of Social Research methodology 5*(3), 215–228.

Jorgensen, Danny L 1989. *Participant Observation: A Methodology for Human Studies*, Volume 15. SAGE.

Kleinman, Sherryl, Barbara Stenross, and Martha McMahon 1994. Privileging fieldwork over interviews: Consequences for identity and practice. *Symbolic Interaction 17*(1), 37–50.

Lamont, Michèle and Virág Molnár 2002. The study of boundaries in the social sciences. *Annual Review of Sociology 28*(1), 167–195.

Laumann, Edward O., Peter V. Marsden, and David Prensky 1989. The Boundary Specification Problem in Network Analysis. In *Research Methods in Social Network Analysis*, pp. 61–87. Routledge.

Lee, Yew-Jin and Wolff-Michael Roth 2004. Making a scientist: Discursive "doing"' of identity and self-presentation during research interviews. *Forum: Qualitative Sozialforschung 5*(1).

McCurdy, Patrick and Julie Uldam 2014. Connecting participant observation positions: Toward a reflexive framework for studying social movements. *Field Methods 26*(1), 40–55.

Merrin, William 2019. President Troll: Trump, 4Chan and Memetic Warfare. In *Trump's media war*, pp. 201–226. Springer.

Mills, C Wright 1959. *The sociological imagination*. Oxford University Press.

Monroe, Burt L., Michael P. Colaresi, and Kevin M. Quinn 2008. Fightin' Words: Lexical Feature Selection and Evaluation for Identifying the Content of Political Conflict. *Political Analysis 16*(4), 372–403.

Nelson, Laura K. 2020. Computational grounded theory: A methodological framework. *Sociological Methods & Research 49*(1), 3–42.

Peel, Leto, Daniel B. Larremore, and Aaron Clauset 2017. The ground truth about metadata and community detection in networks. *Science Advances 3*(5), e1602548.

Strauss, Anselm and Juliet M Corbin 1997. *Grounded theory in practice*. SAGE Publications.

Tavory, Iddo 2010. Of yarmulkes and categories: Delegating boundaries and the phenomenology of interactional expectation. *Theory and Society 39*(1), 49–68.

Watson, David R. and Thomas S. Weinberg 1982. Interviews and the interactional construction of accounts of homosexual identity. *Social Analysis: The International Journal of Social and Cultural Practice 11*, 56–78.