Quantitative Discovery from Qualitative Information: A General-Purpose Document Clustering Methodology*

Justin Grimmer† Gary King‡

October 5, 2009

Abstract

Many people attempt to discover useful information by reading large quantities of unstructured text, but because of known human limitations even experts are ill-suited to succeed at this task. This difficulty has inspired the creation of numerous automated cluster analysis methods to aid discovery. We address two problems that plague this literature. First, the optimal use of any one of these methods requires that it be applied only to a specific substantive area, but the best area for each method is rarely discussed and usually unknowable ex ante. We tackle this problem with mathematical, statistical, and visualization tools that define a search space built from the solutions to all previously proposed cluster analysis methods (and any qualitative approaches one has time to include) and enable a user to explore it and quickly identify useful information. Second, in part because of the nature of unsupervised learning problems, cluster analysis methods are rarely evaluated in ways that make them vulnerable to being proven suboptimal or less than useful in specific data types. We therefore propose new experimental designs for evaluating these methods. With such evaluation designs, we demonstrate that our computer-assisted approach facilitates more efficient and insightful discovery of useful information than either expert human coders or existing automated methods. We (will) make available an easy-to-use software package that implements all our suggestions.

*For helpful advice, coding, comments, or data we thank John Ahlquist, Jennifer Bachner, Jon Bischof, Matt Blackwell, Heidi Brockman, Jack Buckley, Jacqueline Chattopdhyay, Patrick Egan, Adam Glynn, Emily Hickey, Chase Harrison, Dan Hopkins, Grace Kim, Elena Llaudet, Katie Levine, Elena Llaudet, Scott Moser, Jim Pitman, Matthew Platt, Ellie Powell, Maya Sen, Arthur Spirling, Brandon Stewart, and Miya Woolfalk. This paper describes material that is patent pending.

†Ph.D. candidate, Department of Government; Institute for Quantitative Social Science, 1737 Cambridge Street, Harvard University, Cambridge MA 02138; http://www.people.fas.harvard.edu/∼jgrimmer/, jgrimmer@fas.harvard.edu, (617) 710-6803.

‡Albert J. Weatherhead III University Professor, Institute for Quantitative Social Science, 1737 Cambridge Street, Harvard University, Cambridge MA 02138; http://GKing.harvard.edu, king@harvard.edu, (617) 495-2027.

1 Introduction

Creating categories and classifying objects in the categories "is arguably one of the most central and generic of all our conceptual exercises. It is the foundation not only for conceptualization, language, and speech, but also for mathematics, statistics, and data analysis in general. Without classification, there could be no advanced conceptualization, reasoning, language, data analysis or, for that matter, social science research" (Bailey, 1994). Social scientists and most others frequently create categorization schemes, each of which is a lens, or model, through which we view the world (eat this, not that; the bad guys v. good guys; Protestant, Catholic, Jewish, Muslim; Democracy, Autocracy; strongly disagree, disagree, neutral, agree, strongly agree; etc.). Like models in general, any one categorization scheme is never true or false, but it can be more or less useful. Human beings seem to be able to create categorizations instinctively, on the fly, and without the necessity of much explicit thought. Professional social scientists often think carefully when creating categorizations, but commonly used approaches are ad hoc, intuitive, and informal and do not include systematic comparision, evaluation, or verification that their categorizations are in any sense optimal or more useful than other possiblities.

Unfortunately, as we explain below, the limited working memories and computational abilities of human beings means that even the best of us are ill-suited to creating the most useful or informative classification schemes. The paradox, that human beings are unable to use optimally what may be human kind's most important methodological innovation, has motivated a large "unsupervised learning" literature in statistics and computer science to aid human efforts by trying to develop universally applicable cluster analysis methods. In this paper, we address two problems in this important literature.

First, all existing cluster analysis methods are designed for only a limited range of specific substantive problems, but determining which method is most appropriate in any data set is largely unexplained in the literature and difficult or impossible in most cases to determine. Those using this technology are thus left with many options but no guide to the methods or any way to produce such a guide. We attack this problem by developing an easy-to-use method of clustering and discovering new information. Our single approach encompasses all existing automated cluster analysis methods, numerous novel ones we create, and any others a researcher creates by hand; moreover, unlike any existing approach, it is applicable across the vast majority of substantive problems.

Second, the literature offers few if any satisfactory procedures for evaluating categorization schemes or the methods that produce them. Unlike in supervised learning methods or classical statistical estimation, straightforward concepts like unbiasedness or consistency do not apply to the problem of creating categories to derive useful insights. We respond to this somewhat ill-defined but crucial challenge by offering a design for conducting empirical evaluation experiments that reveal the quality of the results and the degree of useful information discovered. We implement these experimental designs in a variety of data sets and show that our computer-assisted human clustering methods dominate what substance matter experts can do alone or how existing automated methods perform, regardless of the content of the data.

Although our methods apply to categories of any type of object, we apply them here to clustering documents containing unstructured text. The spectacular growth in the production and availability of text makes this application of crucial importance throughout the social sciences. Examples include scholarly literatures, news stories, medical records, legislation, blog posts, comments, product reviews, emails, social media updates, audio-to-text summaries of speeches, etc. In fact, emails alone produce a quantity of text equivalent to that in Library of Congress every 10 minutes (Abelson, Ledeen and Lewis, 2008).

Our ultimate goal is to aid "discovery", the process of revealing useful information. Our version of discovery requires three steps: (1) *conceptualization* (as represented in a categorization scheme), (2) *measurement* through this lens (such as classifying documents into the categories), and (3) *verification* of some hypothesis in a way that could be proven wrong. Quantitative approaches tend to focus on measurement and verification, assuming the existence, but not establishing the usefulness, of their categorization scheme. Qualitative research tends to focus on and iterate between conceptualization and measurement, which often results in ignoring or underplaying verification. Existing cluster analysis methods include diverse algorithms for creating categorizations, but without a guide to when they apply or a useful evaluation strategy. We seek to aid all these approaches by improving categorization and the resulting conceptualization in a way that contributes to the whole discovery process, including conceptualization, measurement, and (as we show in our examples) verification of given hypotheses. The point is not to replace humans with atheoretical computer algorithms, but rather to provide *computer-assisted* techniques that improve human performance beyond what anyone could accomplish on their own.

2 The Problem with Clustering

Our specific goal is the discovery of useful information from large numbers of documents, each containing unstructured text. Our starting point in building an approach to discovery is the long tradition in qualitative research that creates data-driven typologies: partitions of the documents into categories that share salient attributes, each partition, or clustering, representing some insight or useful information (Lazardsfeld and Barton, 1965; Bailey, 1994; Elman, 2005). While this is a simple goal to state, and seemingly easy to do by hand, we show here that optimally creating typologies is almost unfathomably difficult. Indeed, humans are ill-suited to make the comparisons necessary to use qualitative methods optimally (Section 2.1), and computer-based cluster analysis algorithms largely developed for other purposes (such as to ease information retrieval or to present search results; see Manning, Raghavan and Schütze, 2008, p.323) face different but similarly insurmountable challenges for our goal (Section 2.2). Recognizing the severe limitations of either humans or computers working alone, we propose a way to exploit the strengths of both the qualitative and quantitative literatures via a new method of computer-assisted human clustering (Section 3).

2.1 Why Johnny Can't Classify

Classifying documents in an optimal way is an extremely challenging computational task that no human being can come close to optimizing by hand.[1] Define a *clustering* as a partition of a set of documents into a set of clusters. Then the overall task involves choosing the "best" (by some definition) among all possible clusterings of a set of n objects (which mathematically is known as the Bell number; e.g., Spivey 2008). The task may sound simple, but merely enumerating the possibilities is essentially impossible for even moderate numbers of documents. Although the number of partitions of two documents is only two (both in the same cluster or each in separate clusters), and the number of partitions of three documents is five, the number of partitions increases

[1]Throughout, we assume that tasks are well defined (with non-trivial functions used to evaluate partitions of documents) or that the problem could be well defined with non-trival functions that encode our evaluations of the partitions. We therefore refer to "optimality" with respect to a specific function used to evaluate the partitions.

very fast thereafter. For example, the number of partitions of a set of 100 documents is 4.76e+115, which is approximately 10^{28} times the number of elementary particles in the universe. Even if we fix the number of partitions, the number is still far beyond human abilities; for example, the number ways of classifying 100 documents into 2 categories is 6.33e+29, or approximately the number of atoms needed to construct more than 10 cars.

Of course, the task of optimal classification involves more than enumeration. Qualitative scholars, recognizing the complexity of clustering documents directly have suggested a number of useful heuristics when creating typologies by hand, but these still require humans to search over an unrealistically large choice set (Elman, 2005). For example, consider a heuristic due to Bailey (1994) which has the analyst beginning with the most fine-grained typology possible and then grouping together the two most closest categories, and then the two next most closest, etc., until a reasonably comprehensible categorization scheme is chosen. This seems like a reasonble procedure, but consider an application to an example he raises, which is a set of items characterized by only 10 dichotomous variables, leading to an initial typology of $2^{10} = 1,024$ categories. However, to compress only the closest two of these categories together would require evaluating more than $4,600$ potential grouping possibilities. This contrasts with a number somewhere between 4 and 7 (or somewhat more if ordered hierarchically) for how many items a human being can keep in short-term working memory (Miller, 1956; Cowan, 2000). Even if this step could somehow be accomplished, it would have to be performed over half a million times in order to reach a number of categories useful for human understanding. Various heuristics have been suggested to simplify this process further, but each is extremely onerous and likely to lead to satisficing rather than optimizing (Simon, 1957). Moreover, inter-coder reliability for tasks like these, even for well-trained coders, is quite low, especially for tasks with many categories (Mikhaylov, Laver and Benoit, 2008).

Finding interesting typologies by hand can of course be done, but no human is capable of doing it well. In short, optimal clustering by hand is infeasible.

2.2 Why HAL Can't Classify

Unfortunately, not only can't Johnny classify, but an impossibly fast computer can't do it either, at least not without knowing a lot about the substance of the problem to which a particular method is applied. That is, the implicit goal of the literature — developing a cluster analysis method that works well across applications — is actually known to be impossible. The "ugly duckling theorem" proves that, without some substantive assumptions, every pair of documents are equally similar and as a result every partition of documents is equally similar (Watanabe, 1969); and the "no free lunch theorem" proves that every possible clustering method performs equally well on average over all possible substantive applications (Wolpert and Macready, 1997; Ho and Pepyne, 2002). Thus, any single cluster analysis method can be optimal only with respect to some specific set of substantive problems and type of data set.

Although application-independent clustering is impossible, very little is known about which substantive problems existing cluster analysis methods work best for. Each of the numerous cluster analysis methods is well defined from a statistical, computational, data analysis, machine learning, or other perspective, but very few are justified in a way that makes it possible to know ex ante in which data set any one would work well. To take a specific example, if one had a corpus of all blog posts about national candidates during the 2008 U.S. presidential primary season, should you use model-based approaches, subspace clustering methods, spectral approaches, grid-based methods, graph-based methods, fuzzy k-modes, affinity propagation, self-organizing maps, or something else? All these and many other proposed clustering algorithms are clearly described in the literature, and many have been implemented in available computer code, but very few hints exist about when any of these methods would work best, well, or better than some other method (Gan, Ma and Wu, 2007).

Consider for example, the finite normal mixture clustering model, which is a particularly "principled statistical approach" (Fraley and Raftery, 2002, p.611). This model is easy to understand, has a well-defined likelihood, can be interpreted from a frequentist or Bayesian perspective, and has been extended in a variety of ways. However, we know from the ugly duckling and no free lunch theorems that no one approach, including this one, is universally applicable or optimal across applications. Yet, we were unable to find any suggestion in the literature about whether a particular corpus, composed of documents of particular substantive topics is likely to reveal its secrets best when analyzed as a mixture of normal densities. The method has been applied to various data sets, but it is seemingly impossible to know when it will work before looking at the results in any application. Moreover, finite normal mixtures are among the simplest and, from a statistical perspective, most transparent cluster analysis approaches available; knowing when most other approaches will work will likely be even more difficult.

Developing intuition for when existing cluster analysis methods work surely must be possible in some special cases, but doing so for most of the rich diversity of available approaches seems infeasible. The most advanced literature along these lines seems to be the new approaches that build time, space, or other structural features of specific data types into cluster analysis methods (Teh et al., 2006; Quinn et al., 2006; Grimmer, 2009). Our approach is useful for discovering structural features; it can also be adapted to include these features, which is useful since the critique in this section applies within the set of all methods that represent structure in some specific way. Of course, the problem of knowing when a cluster analysis method applies occurs in unsupervised learning problems almost by definition, since the goal of the analysis is to discover unknown facts. If we knew ex ante something as specific as the model from which the data were generated up to some unknown parameters (say), we would likely not be at the early discovery stage of analysis.

3 General Purpose Computer-Assisted Clustering

In this section, we introduce a new approach designed to connect methods and substance in unsupervised clustering problems. We do this by combining the output of a wide range of clustering methods with human judgment about the substance of problems judged directly from the results, along with an automated visualization to make this task easier. Our approach breaks the problems posed by the ugly duckling and no free lunch theorems by establishing deep connections between methods and substance — but only after the fact.[2] This means our approach can greatly facilitate discovery (in ways that we show are verifiably better from an empirical standpoint), but at the cost of not being able to use it in isolation to confirm a pre-existing hypothesis, because the users of our method make the final determination about which partition is useful for a particular research

[2]The idea that some concepts are far easier defined by example is commonly recognized in many areas of inquiry. It underlies Justice Potter Stewart's famous threshold for determining obscenity (in *Jacobellis v. Ohio* 378 U.S. 184 (1964)) that "I shall not today attempt further to define the kinds of material I understand to be embraced within that shorthand description; and perhaps I could never succeed in intelligibly doing so. But I know it when I see it. . . ." The same logic been formalized for characterizing difficult-to-define concepts via anchoring vignettes (King et al., 2004).

question. (Section 4.3 illustrates how to verify or falsify hypotheses derived through our approach.) By definition, using such an approach for discovery, as compared to estimation (or confirmation), does not risk bias, since the definition of bias requires a known quantity of interest defined ex ante.

Letting substance enter into the analysis after the fact can be achieved in principle by presenting an extremely long list of clusterings (ideally, all of them), and letting the researcher choose the best one for his or her substantive purposes, such as that which produces the most insightful information or discovery. However, human beings do not have the patience, attention span, memory, or cognitive capacity to evaluate so many clusterings in haphazard order. Our key contribution, then, is to organize these clusterings so a researcher can quickly (usually in 10-15 minutes) select the best one, to satisfy their own substantive objective. We do this by representing each clustering as a point in a two-dimensional visual space, such that clusterings (points) close together in the space are almost the same, and those farther apart may warrent a closer look because they differ in some important way. In effect, this visualization translates the unintepretable chaos of huge numbers of possible clusterings into a simple framework that (we show) human researchers are able to easily comprehend and use to efficiently select one or a small number of clusterings which conveys the most useful information.

To create this space of clusterings, we follow six steps, outlined here and detailed below. First, we translate textual documents to a numerical data set (Section 3.1). (This step is necessary only when the items to be clustered are text documents; all our methods would apply without this step to preexisting numerical data.) Second, we apply (essentially) all clustering methods proposed in the literature, one at a time, to the numerical data set (Section 3.2). Each approach represents different substantive assumptions that are difficult to express before their application, but the effects of each set of assumptions is easily seen in the resulting clusters, which is the metric of most interest to applied researchers. (A new package we have written makes this relatively fast.) Third, we develop a metric to measure the similarity between any pair of clusterings (Section 3.3). Fourth, we use this metric to create a metric space of clusterings, along with a lower dimensional Euclidean representation useful for visualization (Section 3.4).

Fifth, we introduce a "local cluster ensemble" method (Section 3.5) as a way to summarize any point in the space, including points for which there exist no prior clustering methods — in which case they are formed as local weighted combinations of existing methods, with weights based on how far each existing clustering is from the chosen point. This allows for the fast exploration of the space, ensuring that users of the software are able to quickly identify partitions useful for their particular research question. Sixth and finally, we develop a new type of animated visualization which uses the local cluster ensemble approach to explore the metric space of clusterings by moving around it while one clustering slowly morphs into others (Section 3.6), again to rapidly allow users to easily identify the partition (or partitions) useful for a particular research question. We also introduce an optional addition to our method which creates new clusterings (Section 3.7).

3.1 Standard Preprocessing: Text to Numbers

We begin with a set of text documents of variable length. For each, we adopt the most common procedures for representing them quantitatively: we transform to lower case, remove punctuation, replace words with their stems, and drop words appearing in fewer than 1% or more than 99% of documents. For English documents, about 3,500 unique word stems usually remain in the entire corpora. We then code each document with a set of (about 3,500) variables, each coding the number of times a word stem is used in that document.

Although these are very common procedures in the natural language processing literature, they seem surprising given all the information discarded — most strikingly the order of the words (Manning, Raghavan and Schütze, 2008). Since many other ways of reducing text to a numerical data set could be used in our approach (instead, or in addition since we allow multiple representations of the same text), we pause to follow the empirical spirit of this paper by offering an empirical evaluation of this simple approoach. This evaluation also illustrates how easy it is for machines to outdistance human judgment — even as judged by the same human beings.

We begin with 100 diverse essays, each less than about 1,000 words and written by a political scientist. Each gives one view of "the future of political science" published in a book by the same name (King, Schlozman and Nie, 2009). During copyediting, the publisher asked the editors to add to the end of each essay a "see also" with a reference to the substantively closest essay among the remaining 99. The editors agreed to let us run the following experiment to evaluate our coding procedures.

Pairs from	Overall Mean	Evaluator 1	Evaluator 2
Machine	2.24	2.08	2.40
Hand-Coding	2.06	1.88	2.24
Hand-Coded Clusters	1.58	1.48	1.68
Random Selection	1.38	1.16	1.60

Table 1: Evaluators' Rate Machine Choices Better Than Their Own. This table show how preprocessing steps accurately encode information about the content of documents and how automated approaches outperform hand coding efforts.

We begin with a random selection of 25 of the essays. For each, we choose the closest essay among the remaining 99 in the entire book in four separate ways. First, we use our procedure, measuring substantive content solely from each document's numerical summary (and using the cosine as a measure of association between each pair of documents). Second, we asked a group of graduate students and a faculty member at one university to choose, for each of 25 randomly selected essays, the closest other essay they could find among the remaining 99. Third, we asked a second set of graduate students and a faculty member at a different university to group together the essays into clusters of their choosing, and then we created pairs by randomly selecting essays within their clusters. And finally, as a baseline, we matched each essay with a randomly selected essay. We then randomly permuted the order of the resulting $25 \times 4 = 100$ pairs of documents, blinded information about which method was used to create each pair, and sent them back to one graduate student from each group to evaluate. We asked each graduate student to score each pair separately as (1) unrelated, (2) loosely related, or (3) closely related. (Our extensive pretesting indicated that inter-coder reliability would suffer with more categories, but coders are able to understand and use effectively this simple coding scheme.)

Using the same evaluators likely bias the results against our automatically created pairs, but the machine triumphs anyway. Table 1 presents the results, with values indicating the average on this scale from 1 to 3. The first row gives the machine-created pairs which are closest among all methods; it is consistently higher the two control groups created by humans, which in turn is higher than the randomly constructed pairs.

Our evaluators thus judge the automatically created pairs to be more closely related than the pairs they themselves created. This is a small experiment, but it confirms the validity of the standard approach to turning text documents into textual summaries, as well as the point made above that machines can easily outdistance humans in extracting information from large quantities of text.

Our general procedure also accommodates multiple representations of the same documents. These might include tf-idf or other term weighting representations, part of speech tagging, replacing "do" and "not" with "do_not", etc. (Monroe, Colaresi and Quinn, 2008). Likewise, the many variants of kernel methods — procedures to produce a similarity metric between documents without explicitly representing the words in a matrix — could also be included (Shawe-Taylor and Cristianini, 2004).

3.2 Available Clustering Methods

The second step in our strategy involves applying a large number of clustering methods, one at a time, to the numerical representation of our documents. To do this, we have written an R package that can run (with a common syntax) every published clustering method we could find that has been applied to text and used in at least one article by an author other than its developer; we have also included many clustering methods that have not been applied to text before. We developed computationally efficient implementations for the methods included in our program (including variational approximations for the Bayesian statistical methods) so that we can run all the methods on a moderate sized data set in only about 15 minutes; new methods can easily be added to the package as well. Although inferences from our method are typically not affected much, and almost never discontinuously, by including any additional individual method, we still recommend including as many as are available.

To give an idea of the scope of the methods included in our program, we now summarize the types of different clustering algorithms included in our software. Existing algorithms are most often described as either statistical and algorithmic. The statistical models are primarily mixture models, including a large variety of finite mixture models (Fraley and Raftery, 2002; Banerjee et al., 2005; Quinn et al., 2006), infinite mixture models based on the Dirichlet process prior (Blei and Jordan, 2006), and mixture models that cluster both words and documents simultaneously (Blei, Ng and Jordan, 2003). The algorithmic approaches include methods that partition the documents directly, those that create a hierarchy of clusterings, and those which add an additional step to the clustering procedure. The methods include some which identify an exemplar document for each cluster (Kaufman and Rousseeuw, 1990; Frey and Dueck, 2007) and those which do not (Schrodt and Gerner, 1997; Shi and Malik, 2000; Ng, Jordan and Weiss, 2002; von Luxburg, 2007). The hierarchical methods can be further sub-divided into agglomerative (Hastie, Tibshirani and Friedman, 2001), divisive (Kaufman and Rousseeuw, 1990), and other hybrid methods (Gan, Ma and Wu, 2007). To use in our program, we obtain a flat partition of the documents from hierachical clustering created. A final group include methods which group words and documents together simulatenously (Dhillon, 2003) and those which embed the documents into lower dimensional space and then cluster (Kohonen, 2001). Some methods implicitly define a distance metric among documents but, for those that do not, we include many ways to measure the similarity between pairs documents, which is an input to a subset of the clustering methods used here. These include standard measures of distance (Manhattan, Euclidean), angular based measures of similarity (cosine), and many others.

A complete list of the included similarity metrics and clustering methods we have used in our work is available in the software that accompanies this paper. But it is important to recognize that our methodology encompasses any clustering, however created. This includes machine-based categorizations and those created by qualitative heuristics, flat or hierarchical or non-hierarchically grouped, or soft or hard. It can include multiple clusterings for the same methods with different tuning parameters, alternative proximity measures among documents, or any other variation. It can even include clusterings or typologies created by any other procedure, including purely by hand. The only requirement is that each "method" form a proper clustering, with each document assigned either to a single cluster or to different clusters with weights that sum to 1.

3.3 Distance Between Clusterings

We now derive a distance metric for measuring how similar one clustering is to another. We give a qualitative overview here and present mathematical details of our assumptions in Appendix A. We develop our metric axiomatically by stating three axioms that narrow the range of possible choices of distance metrics to only one.

First, the distance between clusterings is a function of the number of pairs of documents not placed together (i.e., in the same cluster) in both clusterings. (We also prove in the appendix that focusing on pairwise disagreements between clusterings is sufficient to encompass differences based on all possible larger subsets of documents, such as triples, quadruples, etc.) Second, we require that the distance be invariant to the number of documents, given any fixed number of clusters in each clustering. Third, we set the scale of the measure by fixing the minimum distance to zero and the maximum distance to $\log(k)$.

As we prove in the Appendix A, only one measure of distance satisfies all three of these reasonable axioms, the *variation of information* (See Equation A.5). This measure has also been derived for different purposes from a larger number of different first principles by Meila (2007).

3.4 The Space of Clusterings

The matrix of distances between each pair of K clusterings can be represented in a K dimensional metric space. In order to visualize this space, we project it down to two Euclidean dimensions. As projection entails the loss of information, the key is to choose a multidimensional scaling method that retains the most crucial information. For our purposes, we need to preserve small distances most accurately, as they reflect clusterings to be combined (in the next section) into *local* cluster ensembles. As the distance between two clusterings increases, a higher level of distortion will affect our results less. This leads naturally to the Sammon multidimensional scaling algorithm (Sammon, 1969); Appendix C defines this algorithm and explains how it satisfies our criteria.

An illustration of this space is given in the central panel of Figure 1, with individual clusterings labeled (we discuss this figure in more detail below).

3.5 Local Cluster Ensembles

In this section we describe how to create a local cluster ensemble, which facilitates the fast exploration of the space of partitions generated in the previous section. A "cluster ensemble" is a technique used to produce a single clustering by averaging in a specific way across many individual clusterings (Strehl and Grosh, 2002; Fern and Brodley, 2003; Law, Topchy and Jain, 2004; Caruana et al., 2006; Gionis, Mannila and Tsaparas, 2005; Topchy, Jain and Punch, 2003). This approach has the advantage of creating a new, potentially better, clustering, but by definition it eliminates the underlying diversity of individual clusterings and so does not work for our purposes. A related technique that is sometimes described by the same term organizes results by performing a "meta-clustering" of the individual clusterings. This alternative procedure has the advantage of preserving some of the diversity of the clustering solutions and letting the user choose, but since no method is offered to summarize the many clusterings within each "meta-cluster," it does not solve our problem. Moreover, for our purposes, the technique suffers from a problem of infinite regress. That is, since any individual clustering method can be used to cluster the clusterings, a researcher would have to use them all to avoid eliminating meaningful diversity in the set of clusterings to be explored. So whether the diversity of clusterings is eliminated by arbitrary choice of meta-clustering method rather than a substantive choice, or we are left with more solutions than we started with, these techniques although useful for some other purposes do not solve our particular problem.

Thus, to preserve diversity — necessary to explore the space of clusterings — and avoid the infinite regress resulting from clustering a set of clusterings, we develop here a method of generating *local cluster ensembles*, which we define as a new clustering created at a point in the space of clusterings from a weighted average of nearby existing clusterings. This approach allows us to preserve the diversity of the individual clusterings while still generating new clusterings that average the insights of many different, but similar, methods. Local cluster ensembles will form the core of our visualization procedure described in the next section.

The procedure requires three steps. First, we define the weights around a user selected point in the space. Suppose that someone applying our software selects $\boldsymbol{x}^* = (x_1^*, x_2^*)$ as the point in our space of clusterings to explore (and therefore the point around which we want to build a local cluster ensemble). The new clustering defined at this point is a weighted average of nearby clusterings with one weight for each existing clustering in the space, so that the closer the existing clustering, the higher the weight. We base the weight for each existing clustering j on a normalized kernel, as $w_j = p(\boldsymbol{x}^*, \sigma^2)/\sum_{m=1}^{J} p(\boldsymbol{x}_m, \sigma^2)$, where $p(\boldsymbol{x}^*, \sigma^2)$ is the height of the kernel (such as a normal or Epanechnikov density) with mean $\boldsymbol{x}^*$ and smoothing parameter σ^2. The collection of weights for all J clusterings is then $\boldsymbol{w} = (w_1, \ldots, w_K)$.

Second, given the weights, we create a similarity matrix for the local cluster ensemble using a voting approach, where each clustering casts a weighted vote for whether each pair of documents appears together in a cluster in the new clustering. First, for a corpora with N documents clustered by method j into K_j clusters, we define an $N \times K_j$ matrix $\boldsymbol{c}_j$ which records how each document is allocated into (or among) the clusters (i.e., so that each row sums to 1). We then horizontally concatenate the clusterings created from all J methods into an $N \times K$ weighted voting matrix $\boldsymbol{V}(\boldsymbol{w}) = \{w_1\boldsymbol{c}_1, \ldots, w_J\boldsymbol{c}_J\}$ (where $K = \sum_{j=1}^{J} K_j$). The results of the election is a new similarity matrix, which we create as $\boldsymbol{S}(\boldsymbol{w}) = \boldsymbol{V}(\boldsymbol{w})\boldsymbol{V}(\boldsymbol{w})'$. This calculation places priority on those cluster analysis methods closest in the space of clusters.

Finally, we create a new clustering for point $\boldsymbol{x}^*$ in the space by applying any coherent clustering algorithm to this new averaged similarity matrix (with the number of clusters fixed to a weighted average of the number of clusters from nearby clusterings, using the same weights). As Appendix D demonstrates, our definition of the local cluster ensemble approach is invariant to the particular choice of clustering method applied to the new averaged similarity matrix. This invariance thus also eliminates the infinite regress problem by turning a meta-cluster method selection problem into a weight selection problem (with weights that are variable in the method). The appendix also shows how our local cluster ensemble approach is closely related to our underlying distance metric defined in Section 3.3. The local cluster ensemble approach will approximate more possible clusterings as additional methods are included, and of course will never be worse, and usually considerably better, in approximating a new clustering than the closest existing observed point.

3.6 Cluster Space Visualization

Figure 1 illustrates our visualization of the space of clusterings, when applied to one simple corpora of documents. This example, which we choose for expository purposes, includes the biographies of each U.S. president from Roosevelt to Obama (see `http://whitehouse.gov`).

The two-dimensional projection of the space of clusterings is illustrated in the figure's central panel, with individual methods labeled. Each method corresponds to one point in this space, and one set of clusters of the given documents. Points corresponding to a labeled method correspond to results from prior research; other points in this space correspond to new clusterings, each constructed as a local cluster ensemble.

A key point is that once the space is constructed, the labeled points corresponding to previous methods deserve no special priority in choosing a final clustering. For example, a researcher should

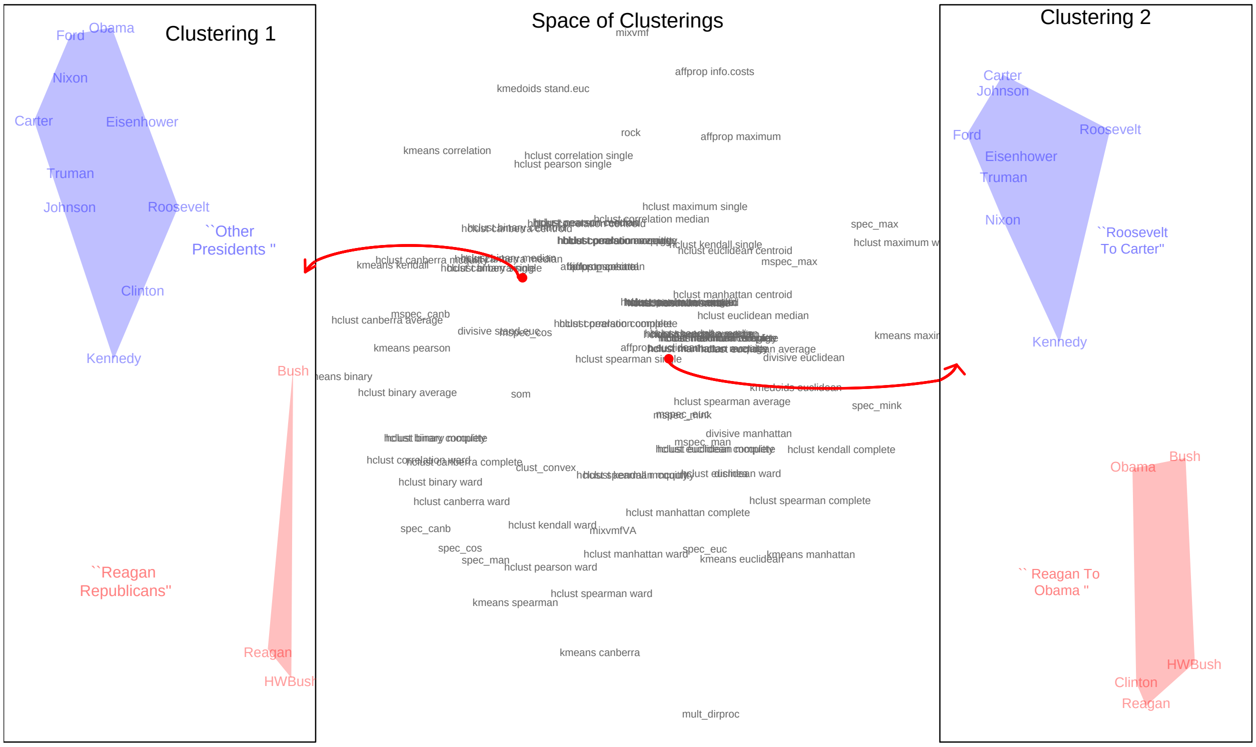

Figure 1: A Clustering Visualization: The center panel gives the space of clusterings, with each name printed representing a clustering generated by that method, and all other points of the space defined by our local cluster ensemble approach that averages nearby clusterings. Two specific clusterings (see red dots with connected arrows), each corresponding to one point in the central space, appear to the left and right; labels in the different color-coded clusters are added for clarification.

not necessarily prefer a clustering from a region of the space with many prior methods as compared to one with few or none. In the end, the choice is the researcher's and should be based on what he or she finds to convey useful information. Since the space itself is crucial, but knowledge of where any prior method exists in the space is not, our visualization software allows the user toggle off these labels so that researchers can focus on clusterings they identify.

The space is formally discrete, since the smallest difference between two clusterings occurs when (for non-fuzzy clustering) exactly one document moves from one cluster to another, but an enormous range of possible clusterings still exists: even this tiny data set of only 13 documents can be partitioned in 27,644,437 possible ways, each representing a different point in this space. A subset of these possible clusterings appear in the figure corresponding to all those clusterings the statistics community has come up with, as well as all possible local cluster ensembles that can be created as weighted averages from them. (The arching shapes in the figure occur regularly in dimension reduction when using methods that emphasize local distances between the points in higher dimensional space; see Diaconis, Goel and Holmes 2008.)

Figure 1 also illustrates two points (as red dots) in the central space, each representing one clustering and portrayed on one side of the central graph, with individual clusters color-coded (and substantive labels added by hand for clarity). Clustering 1, in the left clustering, creates clusters of “Reagan Republicans” (Reagan and the two Bushes) and all others. Clustering 2, on the right, happens to group the presidents into two clusters organized chronologically.

This figure summarizes snapshots of our animated software program at two points. In general, the software is set up so that a researcher can put a single cursor somewhere in the space of clusterings and see the corresponding set of clusters for that point appear in a separate window. The researcher can then move this point and watch the clusters in the separate window morph smoothly from one clustering to another. Our experience in using this visualization often leads us first to check about 4–6 well-separated points, which seems to characterize the main aspects of the diversity of all the clusterings. Then, we narrow the grid further by examining about the same number of clusterings in the local region. Although the visualization offers an enormous number of clusterings, the fact that they are highly ordered in this simple geography makes it possible to understand without much time or effort.

3.7 Optional New Clustering Methods to Add

By encompassing all prior methods, our approach expresses the collective wisdom of the literatures in statistics, biology, computer science, and other areas that have developed cluster analysis methods. By definition, it gives results as good as any individual method included, and better if the much larger space of methods created by our local cluster ensembles produces a useful clustering. Of course, this larger search space is still a small part of the possible clusterings. For most applications, we view this constraint as an advantage, helping to narrow down the enormous “Bell space” of all possible clusterings to a large (indeed larger than has ever before been explored) but yet still managable set of solutions.

However, since the no free lunch theorem applies to subsets of the Bell space of clusterings, there may well be useful insights to be found outside of the space we are exploring. Thus, if, in applying the method, exploring our existing space produces results that do not seem sufficiently insightful, we offer two methods to explore some of the remaining uncharted space.

First, we consider a way of randomly sampling clusterings from the entire Bell space. When desired, a researcher could then add some of these to the original set of clusterings and rerun the same visualization. To do this, we developed a two step method of taking a uniform random draw from the set of all possible clusterings. First, sample the number of clusters K from a multinomial distribution with probability $\text{Stirling}(K, N)/\text{Bell}(N)$ where $\text{Stirling}(K, N)$ is the number of ways to partition N objects into K clusters (i.e., known as the Stirling number of the second kind). Second, conditional on K, obtain a random clustering by sampling the cluster assignment for each document i from a multinomial distribution, with probability $1/K$ for each cluster assignment. If each of the K clusters does not contain at least one document, reject it and take another draw (see Pitman, 1997).

A second approach to expanding the space beyond the existing algorithms directly extends the existing space by drawing larger concentric hulls containing the convex hull of the existing solutions. To do this, we define a Markov chain on the set of partitions, starting with a chain on the boundaries of the existing solutions. To do this, consider a clustering of the data $\boldsymbol{c}_j$. Define $\mathcal{C}(\boldsymbol{c}_j)$ as the set of clusterings that differ by exactly by one document: a clustering $\boldsymbol{c}_j^{'} \in \mathcal{C}(\boldsymbol{c}_j)$ if and only if one document belongs to a different cluster in $\boldsymbol{c}_j^{'}$ than in $\boldsymbol{c}_j$. Our first Markov chain takes a uniform sample from this set of partitions. Therefore, if $\boldsymbol{c}_{j'} \in \mathcal{C}(\boldsymbol{c}_j)$ (and $\boldsymbol{c}_j$ is in the "interior" of the set of partitions) then $p(\boldsymbol{c}_{j'}|\boldsymbol{c}_j) = \frac{1}{NK}$ where N are the number of documents and K is the number of clusters. If $\boldsymbol{c}_{j'} \notin \mathcal{C}(\boldsymbol{c}_j)$ then $p(\boldsymbol{c}_{j'}|\boldsymbol{c}_j) = 0$. To ensure that the Markov chain proceeds outside the existing hull, we add a rejection step: For all $\boldsymbol{c}_{j'} \in \mathcal{C}(\boldsymbol{c}_j)$ $p(\boldsymbol{c}_{j'}|\boldsymbol{c}_j) = \frac{1}{NK}\text{I}(\boldsymbol{c}_{j'} \notin \text{Convex Hull})$. This ensures that the algorithm explores the parts of the Bell space that are not already well described by the included clusterings. To implement this strategy, we use a three stage process applied to each clustering $\boldsymbol{c}_k$: First, we select a cluster to edit with probability $\frac{N_j}{N}$ for each cluster j in clustering $\boldsymbol{c}_k$. Conditional on selecting cluster j we select a document to move with probability $\frac{1}{N_j}$. Then, we move the document to one of the other $K-1$ clusters or to a new cluster, so the document will be sent to a new clustering with probability $\frac{1}{K}$.

4 Evaluating Cluster Analysis Methods

Common approaches to evaluating the performance of cluster analysis methods, which include comparison to internal or supervised learning standards, have known difficulties. Internal standards of comparison define some quantitative measure indicating high similarity of documents within, and low similarity of documents across, clusters. Of course, if this were the goal, it would be possible to define a cluster analysis method with an objective function that optimizes with respect to this measure directly. However, because any one quantitative measure cannot reflect the actual intra-cluster similarity and inter-cluster difference of the substance a researcher happens to be seeking (Section 2.2), "good scores on an internal criterion do not necessarily translate into good effectiveness in an application" (Manning, Raghavan and Schütze, 2008, pp.328–329). The alternative evaluation approach is based on supervised learning standards, which involve comparing the results of a cluster analysis to some "gold standard" set of clusters, pre-chosen by human coders without computer assistance. Although human coders may be capable of assigning documents to a small number of given categories, we know they are incapable of choosing an optimal clustering or one in any sense better than what a computer-assisted method could enable them to create (Section 2.1). As such, using a supervised learning "gold standard" to evaluate an unsupervised learning approach is also of questionable value.

The goal of our analysis — to facilitate discovery — is difficult to formalize mathematically, making a direct evaluation of our method's ability to facilitate discovery complicated. Indeed, some in the statistical literature have even gone so far as to chide those who attempt to use unsupervised learning methods to make systematic discoveries as unscientific (Armstrong, 1967). The primary problem identified in this literature is that existing methods for evaluation do not address the quality of new discoveries.

To respond to these problems, we introduce and implement three new direct approaches to evaluating cluster analysis methods, each addressing a different property of a partition that leads to a good (or useful) discovery. All compare the results of automated methods to human judgment. In each case we use insights from survey research and social psychology to elicit this judgment in ways that people are capable of providing. We first evaluate *cluster quality*, which is the extent to which intracluster similarities outdistance inter-cluster similarities in the substance of a particular application (Section 4.1), followed by what we call *discovery quality*, a direct evaluation by substance matter experts of insights produced by different clusterings in their own data (Section 4.2). In both cases, we show ways of eliciting evaluative information from human coders in a manner that uses their strengths and avoids common human cognitive weaknesses. Third and finally, we offer a substantive application of our method and show how it assists in discovering a specific useful conceptualization and generates new verifiable hypotheses that advance the political science literature (Section 4.3). For this third approach, the judge of the quality of the knowledge learned is the reader of this paper.

4.1 Cluster Quality

We judge cluster quality with respect to a particular corpora by producing a clustering, randomly drawing pairs of documents from the same cluster and from different clusters, and asking human coders to rate the similarity of the documents within each pair (using the same three point scale as in Section 3.1, (1) unrelated, (2) loosely related, (3) closely related), and without conveying how the documents were chosen. Again, we keep our human judges focused on simple tasks they are are able to perform well, in this case comparing only two documents at a time. Our ultimate measure of cluster quality is the average rating of pair similarity within clusters minus the average rating of pair similarity between clusters. (Appendix E introduces a way to save on evaluation costs in measuring cluster quality.)

We apply this measure in each of three different corpora by choosing 25 pairs of documents (13 from the same clusters and 12 from different clusters), computing cluster quality, and averaging over the judgments about the similarity of each pair made separately by many different human coders. We then compare the cluster quality generated by our approach to the cluster quality from a pre-existing hand-coded clustering. What we describe as "our approach" here is a single clustering from the visualization we chose ourselves for this evaluation (we did not participate in evaluating document similarity). This procedure is biased against our method since if we had let the evaluators use our visualization, our approach would have done much better. Although the number of clusters does not necessarily affect the measure of cluster quality, we constrained our method further by requiring it to choose a clustering with approximately the same number of clusters as the pre-existing hand coded clustering.

Press Releases We begin with press releases issued from Senator Frank Lautenberg's Senate office and available on his web site in 24 categories he and his staff chose (`http://lautenberg.senate.gov`). These include appropriations, economy, gun safety, education, tax, social security, veterans, etc. We randomly selected 200 press releases for this experiment. This application represents a high evaluation standard since the documents, the categorization scheme, and the classification of each document into a category were all created by the same group (the Senator and his staff) at great time and expense.

The top line in Figure 2 gives the results for the difference in our method's cluster quality minus the cluster quality from Lautenberg's hand-coded categories. The point estimate appears as a dot, with a thick line for the 80% confidence interval, and thin line for the 95% interval. The results, appearing to the right of the vertical dashed line that marks zero, indicate that our method produces a clustering with unambiguously higher quality than the author of the documents produced by hand. (We give an example of the substantive importance of this result in Section 4.3.)

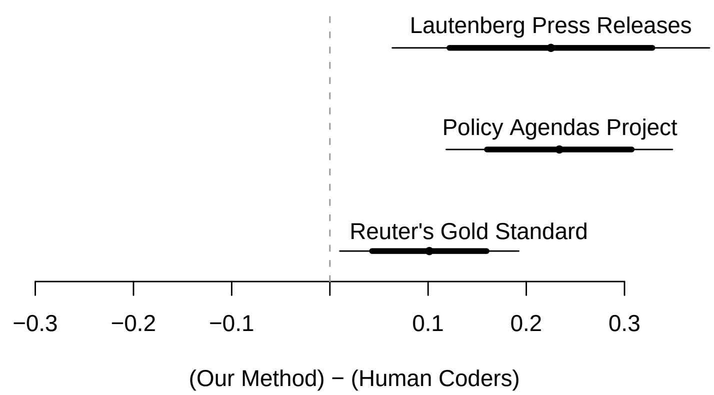

Figure 2: Cluster Quality Experiments: Each line gives a point estimate (dot), 80% confidence interval (dark line), and 95% confidence interval (thin line) for a comparison between our automated cluster analysis method and clusters created by hand. Cluster quality is defined as the average similarity of pairs of documents from the same cluster minus the average similarity of pairs of documents from different clusters, as judged by human coders one pair at a time.

State of the Union Messages Our second example comes from an analysis of all 213 quasi-sentences in President George W. Bush's 2002 State of the Union address, hand coded by the Policy Agendas Project (Jones, Wilkerson and Baumgartner, 2009). Each quasi-sentence (defined in the original text by periods or semicolon separators) takes the role of a document in our discussion. The authors use 19 policy topic-related categories, including agriculture, banking & commerce, civil rights/liberties, defense, education, etc. Quasi-sentences are difficult tests because they are very short and may have meaning obscured by the context, which most automated methods ignore.

The results of our cluster quality evaluation appear as the second line in Figure 2. Again, our automated method clearly dominates that from the hand coding, which can be seen by the whole 95% confidence interval appearing to the right of the vertical dashed line. These results do not imply that anything is wrong with the Policy Agendas classification scheme, only that there seems to be more information in the data they collected than their categories may indicate.

Our exploration of these data also suggests some insights not available through the project's coding scheme. In particular, we found that the largest cluster of statements in Bush's address were those that addressed the 9/11 tragedy, including many devoid of immediate policy implications, and so lumped are into a large "other" category by the project's coding scheme, despite considerable political meaning. For example, "And many have discovered again that even in tragedy, especially in tragedy, God is near." or "We want to be a Nation that serves goals larger than self." This cluster thus conveys how the Bush administration's response to 9/11 was sold rhetorically to resonate with his religious supporters and others, all with considerable policy content. For certain research purposes, this discovery may reflect highly valuable additional information.

Reuters News Stories For a final example, we use 250 documents randomly drawn from the "Reuters-21578" news story categorization. This corpus has often been used as a "gold standard" baseline for evaluating clustering (and supervised learning classification) methods in the computer science literature (Lewis, 1999). In this collection, each Reuters financial news story from 1987 has been classified by the Reuters news organization (with help from a consulting firm) into one of 22 categories, including trade, earnings, copper, gold, coffee, etc. We again apply the same evaluation methodology; the results, which appear as the bottom line in Figure 2, indicates again that our approach has unambiguously higher cluster quality than Reuter's own gold standard classification.

4.2 Discovery Quality

Cluster quality is an essential component of understanding and extracting information from unstructured text, but the ultimate goal of cluster analysis for most social science purposes is the discovery of useful information. Along with each example in the previous section, we reported some evidence that useful discoveries emerged from the application of our methodology, but here we address the question more directly by showing that it leads to more informative discoveries for researchers engaged in real scholarly projects. We emphasize that this is an unusually hard test for a statistical method, and one rarely performed; it would be akin to requiring not merely that a standard statistical method has certain properties like being unbiased, but also, when given to researchers and used in practice, that they actually use it appropriately and estimate their quantities of interest correctly.

The question we ask is whether the computer assistance we provide helps. To perform this evaluation, we recruited two scholars in the process of evaluating large quantities of text in their own work-in-progress, intended for publication (one faculty member, one senior graduate student). We offered an analysis of their text in exchange for their participation in our experiment. One had a collection of documents about immigration in America in 2006; the other was studying a longer period about how genetic testing was covered in the media. (To ensure the right of first publication goes to the authors, we do not describe the specific insights we found here and instead only report how they were judged in comparison to those produced by other methods.) Using the collection of texts from each researcher, we applied our method, the popular k-means clustering methodology (with variable distance metrics), and one of two more recently proposed clustering methodologies — the Dirichlet process prior and the mixture of von Mises Fisher distributions, estimated using both the EM algorithm and a variational approximation. We used two different clusterings from each of the three cluster analysis methods applied in each case. For our method, we again biased the results against our method and this time chose the two clusterings ourselves instead of letting them use our visualization.

We then created an information packet on each of the six clusterings. This included the proportion of documents in each cluster, an exemplar document, and a brief automated summary of the substance of each cluster, using a technique that we developed. To create the summary, we first identified the 10 most informative words stems for each cluster, in each clustering (i.e., those with the highest "mutual information"). The summary then included the full length word most commonly associated with each chosen word stem. We found through much experimentation, that words selected in this way usually provide an excellent summary of the topic of the documents in a cluster.

We then asked each researcher to familiarize themselves with the six clusterings. After about 30 minutes, we asked each to perform all $\binom{6}{2} = 15$ pairwise comparisons between the clusterings and in each case to judge which clustering within a pair is "more informative". We are evaluating two clusterings from each cluster analysis method, and so label them 1 and 2, although the numbers are not intended to convey order. In the end, we want a cluster analysis methodology that produces at least one method that does well. Since the user ultimately will be able to judge and choose among results, having a method that does poorly is not material; the only issue is how good the best one is.

Figure 3 gives a summary of our results, with arrows indicating dominance in pairwise comparisons. In the first (immigration) example, illustrated at the top of the figure, the 15 pairwise comparisons formed a perfect Guttman scale (Guttman, 1950) with "our method 1" being the Condorcet winner (i.e., it beat each of the five other clusterings in separate pairwise comparisons). (This was followed by the two mixtures of Von Mises Fisher distribution clusterings, then "our method 2", and then the two k-means clusterings.) In the genetics example, our researcher's evaluation produced one cycle, and so it was close to but not a perfect Guttman scale; yet, "our method 1" was again the Condorcet winner. (Ranked according to the number of pairwise wins, after "our method 1" was one of the k-means clusterings, then "our method 2", then other k-means clustering, and then the two Dirichlet process cluster analysis methods. The deviation from a Guttman scale occurred among the last three items.)

"Immigration" Discovery Experiment:

Our Method 1 ⟶ vMF VA ⟶ vMF EM ⟶ **Our Method 2** ⟶ K-Means, Cosine ⟶ K-Means, Euc.

"Genetic testing" Discovery Experiment:

Our Method 1 ⟶ **{Our Method 2**, K-Means Max, K-means Canberra**}** ⟶ Dir Proc. 1 ⟶ Dir Proc 2

Figure 3: Results of Discovery Experiments, where $A \rightarrow B$ means that clustering A is judged to be "more informative" than B in a pairwise comparison, {with braces grouping results in the second experiment tied due to an evaluator's cyclic preferences.}. In both experiments, a clustering from our method is judged to beat all others in pairwise comparisons.

4.3 Partisan Taunting: An Illustration of Computer-Assisted Discovery

We now give a brief report of an example of the whole process of analysis and discovery using our approach applied to a real example. We develop a categorization scheme that advances one in the literature, measure the prevalence of each of its categories in a new out-of-sample set of data to show that the category we discovered is common, develop a new hypothesis that occurred to us because of the new lens provided by our new categorization scheme, and then test it in a way that could be proven wrong.

In a famous and monumentally important passage in the study of American politics, Mayhew (1974, p.49ff) argues that "congressmen find it electorally useful to engage in. . . three basic kinds of activities" — credit claiming, advertising, and position taking. This typology has been widely used over the last 35 years, remains a staple in the classroom, and accounts for much of the core of several other subsequently developed categorization schemes (Fiorina, 1989; Eulau and Karps, 1977; Yiannakis, 1982). In the course of preparing our cluster analysis experiments in Section 4.1, we found much evidence for all three of Mayhew's categories in Senator Lautenberg's press releases, but we also made what we view as an interesting new discovery.

We illustrate this discovery process in Figure 4, where the top panel gives the space of clusterings we obtain when applying our methodology to Lautenberg's press releases (i.e., like Figure 1). Recall that each name in the space of clusterings in the top panel corresponds to one clustering obtained by applying the named clustering method to the collection of press releases; any point in the space between labeled points defines a new clustering using our local cluster ensemble approach; and

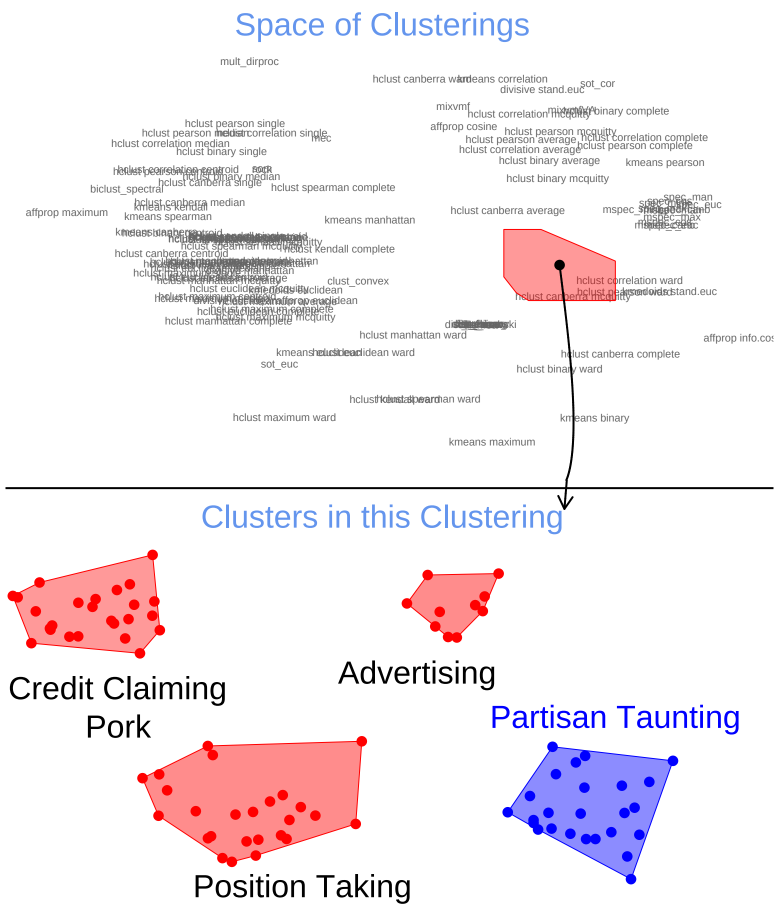

Figure 4: Discovering Partisan Taunting: The top portion of this figure presents the space of clustering solutions of Frank Lautenberg's (D-NY) press releases. Partisan taunting could be easily discovered in any of the clustering solutions in the red region in the top plot. The bottom plot presents the clusters from a representative clustering within the red region at the top (represented by the black dot). Three of the clusters (in red) align with Mayhew's categories, but we also found substantial *partisan taunting* cluster (in blue), with Lautenberg denigrating Republicans in order to claim credit, position take, and advertise. Other points in the space have different clusterings but all clearly reveal the partisan taunting category.

nearby points have clusterings that are more similar than those farther apart.

The clusters within the single clustering represented by the black point in the top panel is illustrated in the bottom panel, with individual clusters comprising Mayhew's categories of claiming credit, advertising, and position taking (all in red), as well as an activity that his typology obscures, and he does not discuss. We call this new category *partisan taunting* (in blue), and describe it below.

Date	Lautenberg Category	Quote
2/19/2004	Civil Rights	“The Intolerance and discrimination from the Bush administration against gay and lesbian Americans is astounding”
2/24/2004	Government Oversight	“Senator Lautenberg Blasts Republicans as ‘Chicken Hawks’ ”
8/12/2004	Government Oversight	“John Kerry had enough conviction to sign up for the military during wartime, unlike the Vice President [Dick Cheney], who had a deep conviction to avoid military service”
12/7/2004	Homeland Security	“Every day the House Republicans dragged this out was a day that made our communities less safe”
7/19/2006	Healthcare	“The scopes trial took place in 1925. Sadly, President Bush’s veto today shows that we haven’t progressed much since then.”

Table 2: Examples of Partisan Taunting in Senator Lautenberg’s Press Releases

Each of the other points in the red region in the top panel represent clusterings that also clearly suggest partisan taunting as an important cluster, although with somewhat different arrangements of the other clusters. That is, the user would only need to examine one point anywhere within this (red) region to have a good chance at discovering partisan taunting as a potentially interesting category.

Examples of partisan taunting appear in Table 2. Unlike any of Mayhew’s categories, each of the colorful examples in the table explicitly reference the opposition party or one of its members, using exaggerated language to put them down or devalue their ideas. Most partisan tauting examples also overlap two or three of Mayhew’s existing categories, which is good evidence of the need for this separate, and heretofore unrecognized, category.

Partisan taunting provides a new category of Congressional speech that emphasizes the interactions inherent between members of a legislature. Mayhews (1974) original theory supposed that members of Congress were atomistic rational actors, concerned only with optimizing their own chance of reelection. Yet, legislators interact with each other regularly, criticizing and supporting ideas, statements, and actions. This interaction is captured with partisan taunting, but absent from the original typology.

Examples from Lautenberg’s press releases and contemporary political discourse suggests new insights into Congressional behavior. Partisan taunting creates the possibilty of *negative* credit claiming: when members of Congress undermine the opposing party’s efforts to claim credit for federal funds. For example, the DCCC issued a press release accusing Mary Bono Mack (R-CA,45) of acting "hypocritically" for announcing "$40 million for two long-awaited improvement projects to I-10, even though she voted against the improvements". Partisan taunting also allows members of a party to claim credit for legislative work even when no reform actually occurred. Both Democrats and Republican caucuses regularly issue statements, blaming inaction in the Congress on the other party. For example a June 27, 2007 press release from the Senate Democratic caucus reads, "Senate Republicans blocked raising the minimum wage".

Partisan taunting is also an important element of position taking, allowing members of Congress to juxtapose their own position against the other party's. Senator Lautenberg used this strategy in a press release when he "filed an amendment to rename the 'Tax Reconciliation Act of 2005,' to reflect the true impact the legislation will have on the nation if allowed to pass. Senator Lautenberg's amendment would change the name of the measure to 'More Tax Breaks for the Rich and More Debt for Our Grandchildren Deficit Expansion Reconciliation Act of 2006.' The Republican bill would provide more tax cuts to the wealthiest Americans while saddling our grandchildren with additional debt."

Partisan taunting also overlaps the category of advertising, which occurs in Lautenberg's press release when he "Expresses Shock Over President Bush's Mock Search for Weapons of Mass Destruction". While devoid of policy content, this statement allows Lautenberg to appear as a sober statesman next to a juvenille administration joke.

Our technique has thus produced a new and potentially useful conceptualization for understanding Senator Lautenberg's 200 press releases. Although asking whether the categorization is "true" makes no sense, this modification to Mayhew's categorization scheme would seem to pass the tests for usefulness given in Section 4.1. We now show that it is also useful for *out-of-sample* descriptive purposes and separately for generating and rigorously testing other hypotheses suggested by this categorization.

We begin with a large out-of-sample test of the descriptive merit of the new category, for which we analyze all 64,033 press releases from all 301 Senators during the three years 2005–2007. To do this, we developed a coding scheme that includes partisan taunting, other types of taunting (to make sure our first category is well defined), and other types of press releases (including Mayhew's three categories). We then randomly selected 500 press releases and had three research assistants assign each press release to a category (resolving any disagreements by reading the press releases ourselves). Finally, we applied the supervised learning approach to text analysis given by (Hopkins and King, 2009, forthcoming) to the entire set of 64,033 press releases to estimate the percent of press releases which were partisan taunts for each senator in each year. (By setting aside a portion of this training set, we verified that the Hopkins-King methodology produced highly accurate estimates in these data.)

Overall, we find that 27% of press releases among these 301 Senator-years were partisan taunts, thus confirming that this category was not merely an idiosyncrasy of Senator Lautenberg. Instead partisan taunting seems to play a central role in the behavior many Senators find it useful to engage in. Indeed, it may even define part of what it means to be a member of the party in government. The histogram in the left panel of Figure 5 gives the distribution of taunting behavior in our data; it conveys the large amount of taunting across numerous press releases, as well as a fairly large disperson across senators and years in taunting behavior.[3]

Finally, analyzing Senator Lautenberg's press releases led us to consider the role of taunting behavior in theories of democratic representation. Almost by definition, partisan taunting is antithetical to open *deliberation* and compromise for the public good (Gutmann and Thompson, 1996). Thus, an important question is who taunts and when — which led us to the hypothesis that taunting would be less likely to occur in competitive senate seats. The idea is that taunting is most effective when a senator has the luxury of preaching to the choir and warning his or her partisans of the opposition (which has few votes); if instead, a politician's electoral constituency is composed of large numbers of opposition party members, we would expect partisan taunting to be less effective and thus less used. If true, this result poses a crucial tension in democratic representation. Deliberation is seen as a normative good, but the degree to which a representative is a *reflection* of his or her constituency is also often seen to be an important component of democracy (Miller and Stokes, 1963; Pitkin, 1972). However, if our hypothesis is empirically correct, then democracies may have a zero sum choice between deliberation, which occurs more often in the absence of partisan

[3]The top 10 Senator-year taunters include Baucus (D-MT), 2005; Byrd (D-WV), 2007; Thune (R-SD), 2006; Ensign (R-NV), 2005; McConnell (R-KY), 2006; Biden (D-DE), 2005; Reid (D-NV), 2005; Coburn (R-OK), 2007; Sarbanes (D-MD), 2006; Kennedy (D-MA), 2007.

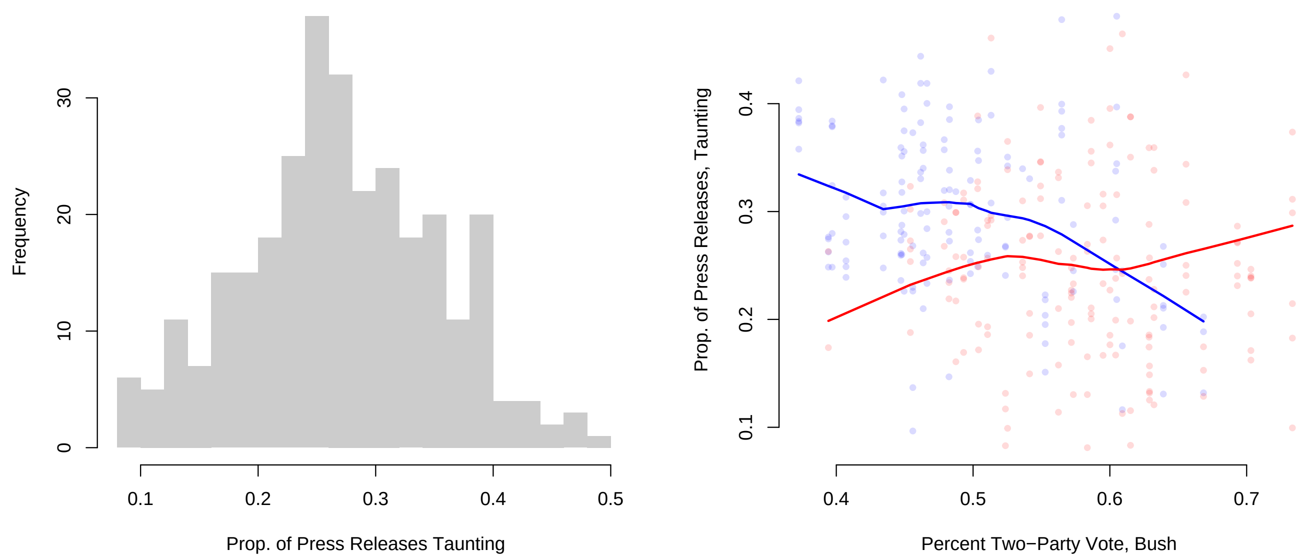

Figure 5: Partisan Taunting Hypothesis Verification. The left panel shows the distribution in partisan taunting in senators' press releases and the right panel demonstrates that taunting is more likely when senators are in less competitive states. Each of the 301 points in the right panel represents the results of an analysis of one year's worth of a single senator's press releases, with blue for Democrats and red for Republicans.

taunting and thus in the most competitive states, and reflection, which by definition occurs in the least competitive states.

By using our large data set of press releases, we construct an out-of-sample test of our hypothesis. The right panel of Figure 5 gives the results. Each dot in this figure represents one senator-year, with red for Republicans and blue for Democrats. The horizontal axis is the proportion of the 2004 two-party vote for George W. Bush — a measure of the size of the underlying Republican coalition in each state, separate from all the idiosyncratic features of individual senatorial campaigns. We also portray the dominant patterns with a smoothed (LOESS) line for the Republicans (in blue) and Democrats (in red). The results overall clearly support the hypothesis: As states become more Republican (moving from left to right), partisan taunting by Republicans increase, whereas partisan taunting by Democrats decline.

Of course, much more can be done with this particular empirical example, which is in fact precisely the point: our clustering methodology helped us choose a new categorization scheme to understand an aspect of the world in a new way, a new concept represented as a new category, a new hypothesis capable of being proven wrong, and a rigorous out-of-sample validation test for both describing and explaining the variation in the prevalence of this category among all Senators.

5 Concluding Remarks

We introduce in this paper a new encompassing approach to unsupervised learning through cluster analysis. We also introduce new empirically based procedures for evaluating this and other cluster analytic methods and their resulting clusterings that use human judgment in a manner consistent with their cognitive strengths. Through a variety of examples, we demonstrate how this approach can relatively easily unearth new discoveries of useful information from large quantities of unstructured text.

Given the ongoing spectacular increase in the production and availability of unstructured text about subjects of interest to political scientists, and the impossibility of assimilating, summarizing, or even characterizing much of it by reading or hand coding, the most important consequence of this research may be its potential for social scientists to help efficiently unlock the secrets this information holds.

For methodologists and statisticians working on developing new methods of cluster analysis, this research also offers new ways of evaluating their products. Research that follows up on our strategy by creating new ways of encompassing existing methods might be designed to make the process easier, visualized in other ways, or computationally faster. Most of the research currently being done is focused on developing individual (i.e., non-encompassing) methods; we know that, by definition, any one individual method cannot outperform the approach proposed here, but new individual methods may be able to improve our approach if included in the cluster methods we encompass. For that purpose, we note that the most useful new individual methods would be those which fill empty areas in the space of clusterings, especially those outside the convex hull of existing methods in this space. Methods that produce clusterings for many data sets close to others would not be as valuable.

Finally, we note that our general approach can logically be extended to continuous variables for methods such as factor analysis or multidimensional scaling. These and other approaches have been both extremely useful in a wide variety of fields and at the same time occasionaly denigrated for their lack of attention to falsifiability and validation (Armstrong, 1967). We thus hope that not only will our specific methods and approach prove useful in building on the advantages of these approaches, but that our general framework for separating discovery from empirical validation, and ensuring the tenets of both are satisfied, will be of use in many types of unsupervised learning problems.

A Defining The Distance Between Clusterings

We now give the mathematical foundation for the results outlined in Section 3.3. Each cluster method j ($j = 1, \ldots, J$) produces a partition (or "clustering") of the documents with K_j clusters assumed (or estimated). Denote by c_{ikj} an indicator of whether (or the extent to which) document i is assigned to cluster k under method j. For "hard" cluster algorithms (those that assign a document to only one cluster), $c_{ikj} \in \{0,1\}$; for "soft" methods, $c_{ikj} \in [0,1]$; and for both $\sum_{k=1}^{K_j} c_{ikj} = 1$ for all k and j. The K_j-vector denoting document i's cluster membership from method j is given by $\boldsymbol{c}_{ij}$ and is an element of the $K_j - 1$ dimensional simplex. Then we characterize a full clustering for method j with the $N \times K_j$ matrix $\boldsymbol{c}_j$.

Our distance metric builds on entropy, a function H that maps from the proportion of documents in each category to a measure of information in the documents. For clustering $\boldsymbol{c}_j$, define the proportion of documents assigned to the k^{th} category as $\sum_{i=1}^{N} \frac{c_{ijk}}{N} = p_j(k)$ and denote as $\boldsymbol{p}_j = (p_j(1), \ldots, p_j(K))$ the vector describing the proportion of documents assigned to each category. The space of all possible proportions for a K-component cluster is then Δ^K, a $K-1$ dimensional simplex. For notational simplicity, we allow $H(\boldsymbol{c}_j)$ and $H(\boldsymbol{p}_j)$ to represent the same quantity. The entropy of a clustering $\boldsymbol{c}_j$ (Mackay, 2003):

$$H(\boldsymbol{c}_j) \equiv H(p_j(1), p_j(2), \ldots, p_j(K)) \quad = \quad -\sum_{k=1}^{K} p_j(k) \log p_j(k). \tag{A.1}$$

We now develop a measure of distance between clusterings based upon a (rescaled) measure of pairwise disagreements. Denote by $d(\boldsymbol{c}_j, \boldsymbol{c}_{j'})$ our candidate measure of the distance between two clusterings, $\boldsymbol{c}_j$ and $\boldsymbol{c}_{j'}$. Define $\text{pair}(\boldsymbol{c}_j, \boldsymbol{c}_{j'})$ as the number of documents in the same cluster in $\boldsymbol{c}_j$ but not in $\boldsymbol{c}_{j'}$ plus the pairs of documents in $\boldsymbol{c}_{j'}$ not in $\boldsymbol{c}_j$. The pair function includes the information in all higher order subsets, such as triples, quadruples, etc. This is well-known, but we offer a simple proof by contradiction here. Suppose, by way of contradiction, that clustering $\boldsymbol{c}_j$ and $\boldsymbol{c}_z$ agree on all pairs, but disagree on some larger subset m. This implies there exists a group of documents $c_{1j}, \ldots, c_{mj}$ grouped in the same cluster in $\boldsymbol{c}_j$ but not grouped in $\boldsymbol{c}_z$. But for this to be true, then there must be at least m pair differences between the two clusterings, contradicting our assumption that there are no pairwise disagreements. Note that the converse is not true: Two clusterings could agree about all subsets of size $m > 2$ but disagree about the pairs of documents that belong together.

We use three assumptions to derive the properties of our distance metric. First, we assume that the distance metric should be based upon the number of pairwise disagreements (encoded in the pair function). We extract two properties of our metric directly from this assumption. First, denote the maximum possible distance between clusterings as that which produces the maximum number of pairwise disagreements about the cluster in which the two documents belong. Denote $c(1, N)$ as the clustering where all N documents are placed into one cluster and $\boldsymbol{c}(N, N)$ the clustering where all N documents are placed into N individual clusters. Then the maximum possible pairwise disagreements is between $\boldsymbol{c}(1, N)$ and $\boldsymbol{c}(N, N)$. (Note that $c(1, N)$ implies $\binom{N}{2}$ pairs, while $c(N, N)$ implies 0 pairs, implying $\binom{N}{2}$ disagreements, the largest possible disagreement.) In addition, for each clustering $\boldsymbol{c}_j$,

$$\text{pair}(\boldsymbol{c}(1, N), \boldsymbol{c}(N, N)) \quad = \quad \text{pair}(\boldsymbol{c}(1, N), \boldsymbol{c}_j) + \text{pair}(\boldsymbol{c}(N, N), \boldsymbol{c}_j). \tag{A.2}$$

Our second property extracted from the focus on pairwise disagreements ensures that partitions with smaller distances are actually more similar — have fewer pairwise disagreements — than other partitions with larger distances. Define the *meet* between two clusterings $\boldsymbol{c}_j$ and $\boldsymbol{c}_k$ as a new (compromise) clustering, denoted $\boldsymbol{c}_j \times \boldsymbol{c}_k$, which assigns pairs of documents to the same cluster if both of the component clusterings agree they belong in the same cluster. If the two clusterings disagree, then the pair of documents are not assigned to the same cluster. A general property of a meet is that it lies "between" two clusterings, or for any clusterings $\boldsymbol{c}_z$ and $\boldsymbol{c}_m$,

$$\text{pair}(\boldsymbol{c}_z, \boldsymbol{c}_m) \quad = \quad \text{pair}(\boldsymbol{c}_z \times \boldsymbol{c}_m, \boldsymbol{c}_z) + \text{pair}(\boldsymbol{c}_z \times \boldsymbol{c}_m, \boldsymbol{c}_m). \tag{A.3}$$

Using the pair function and an additional assumption — invariance to the number of documents included in the clustering — we define a third property of our metric: how the shared information changes as the number of clusters change. Consider the case where we *refine* a clustering $\boldsymbol{c}_j$ by dividing documents in cluster c_{jk} among a set of newly articulated clusters, $\boldsymbol{c}^{'}(n_{jk})$, and where the new clustering is $\boldsymbol{c}^{'}_j$. (If we restrict attention to the n_{jk} documents originally in cluster k in clustering j, then c_{jk} is the clustering that assigns all n_{jk} documents to the same cluster, so we write it as $\boldsymbol{c}(1, n_{jk})$.) A property of the pair function is that,

$$\text{pair}(\boldsymbol{c}_j, \boldsymbol{c}^{'}_j) \;\; = \;\; \sum_{k=1}^{K_j} \text{pair}(\boldsymbol{c}(1, n_{jk}), \boldsymbol{c}^{'}(n_{jk})) \tag{A.4}$$

Using Equation A.4, we apply the invariance assumption to rescale the pair function. Therefore, we require the distance between $\boldsymbol{c}_j$ and $\boldsymbol{c}^{'}_j$ to be $d(\boldsymbol{c}_j, \boldsymbol{c}^{'}_j) = \sum_{k=1}^{K} \frac{n_{jk}}{n} d(\boldsymbol{c}(1, n_{jk}), \boldsymbol{c}^{'}_{jk})$.

The final property employs the pair function plus a scaling axiom to define the maximum distance for a fixed number of clusters K. Call the clustering that places the same number of documents into each cluster $\boldsymbol{c}(\text{uniform}, K)$ (if this clustering exists). Then the clustering with the most pairwise disagreements with $\boldsymbol{c}(\text{uniform}, K)$ is $\boldsymbol{c}(1, N)$ and so bounding on this distance bounds all smaller distances. We use a scaling assumption to require that $d(\boldsymbol{c}(\text{uniform}, K), \boldsymbol{c}(1, N)) = \log K$, i.e., that the distance between an evenly spread out clustering and a clustering that places all documents into the same category increases with the number of categories at a logarithmic rate.

Our three assumptions, and the four properties extracted from these assumptions, narrow the possible metrics to a unique choice: the *variation of information* (VI), based on the shared or conditional entropy between two clusterings Meila (2007). Further, it is a distance metric (even though we made no explicit assumptions that our distance measure be a metric). We define the VI metric by considering the distance between two arbitrary clusterings, $\boldsymbol{c}_j$ and $\boldsymbol{c}^{'}_j$. Define the proportion of documents assigned to cluster k in method j and cluster $k^{'}$ in method $j^{'}$ as $\boldsymbol{p}_{jj'}(k, k^{'}) = \sum_{i=1}^{N} c_{ikj} c_{ik'j'}/N$.

Given the joint-entropy definition of shared information between $\boldsymbol{c}_j$ and $\boldsymbol{c}_{j'}$, $H(\boldsymbol{c}_j, \boldsymbol{c}_{j'}) = -\sum_{k=1}^{K} \sum_{k'=1}^{K} \boldsymbol{p}_{jj'}(k, k^{'}) \log \boldsymbol{p}_{jj'}(k, k^{'})$, we seek to determine the amount of information cluster $\boldsymbol{c}_j$ adds if we have already observed $\boldsymbol{c}_{j'}$. A natural way to measure this additional information is with the conditional entropy, $H(\boldsymbol{c}_j | \boldsymbol{c}_{j'}) = H(\boldsymbol{c}_j, \boldsymbol{c}_{j'}) - H(\boldsymbol{c}_{j'})$, which we make symmetric by adding together the conditional entropies: (Meila, 2007),

$$d(\boldsymbol{c}_j, \boldsymbol{c}_{j'}) \equiv VI(\boldsymbol{c}_j, \boldsymbol{c}_{j'}) \quad = \quad H(\boldsymbol{c}_j|\boldsymbol{c}_{j'}) + H(\boldsymbol{c}_{j'}|\boldsymbol{c}_j). \tag{A.5}$$

B Properties of the Pair Function

In this section we prove various properties of the pair function.

Lemma 1. *For all clusterings* $\boldsymbol{c}_j$, *pair*$(\boldsymbol{c}(1,N), \boldsymbol{c}(N,N)) =$ *pair*$(\boldsymbol{c}(1,N), \boldsymbol{c}_j) +$ *pair*$(\boldsymbol{c}(N,N), \boldsymbol{c}_j)$

Proof. Note that $\boldsymbol{c}(1,N)$ implies $\binom{N}{2}$ pairs of documents, so pair$(\boldsymbol{c}(1,N), \boldsymbol{c}(N,N)) = \binom{N}{2}$. Any $\boldsymbol{c}_j$ will have $g = \sum_{k=1}^{K_j} \binom{n_{jk}}{2}$ pairs of documents, where n_{jk} represents the number of documents assigned to the k^{th} cluster in clustering j. Therefore, pair$(\boldsymbol{c}(1,N), \boldsymbol{c}_j) = \binom{N}{2} - g$. If all clusterings are placed into their own clusters, then there are no pairs of clusters, so pair$(\boldsymbol{c}(N,N), \boldsymbol{c}_j) = g$. Adding these two quantities together we find that, pair$(\boldsymbol{c}(1,N), \boldsymbol{c}_j) +$ pair$(\boldsymbol{c}(N,N), \boldsymbol{c}_j) = \binom{N}{2} =$ pair$(\boldsymbol{c}(1,N), \boldsymbol{c}(N,N))$. So, we require for our distance metric that $d(\boldsymbol{c}(1,N), \boldsymbol{c}(N,N)) = d(\boldsymbol{c}(1,N), \boldsymbol{c}_j) + d(\boldsymbol{c}(N,N), \boldsymbol{c}_j)$ for all possible clusterings $\boldsymbol{c}_j$. □

Lemma 2. *For all clusterings* $\boldsymbol{c}_z$ *and* $\boldsymbol{c}_m$, *pair*$(\boldsymbol{c}_z, \boldsymbol{c}_m) =$ *pair*$(\boldsymbol{c}_z \times \boldsymbol{c}_m, \boldsymbol{c}_z) +$ *pair*$(\boldsymbol{c}_z \times \boldsymbol{c}_m, \boldsymbol{c}_m)$

Proof. Define $g^z = \sum_{k=1}^{K} \binom{n_{zk}}{2}$ and $g^m = \sum_{k=1}^{K'} \binom{n_{mk}}{2}$ and call the number of pairs where the two clusterings agree g^{agree}. Then pair$(\boldsymbol{c}_z, \boldsymbol{c}_m) = (g^z - g^{\text{agree}}) + (g^m - g^{\text{agree}})$. $\boldsymbol{c}_m \times \boldsymbol{c}_z$ places a pair of documents into the same cluster if and only if $\boldsymbol{c}_z$ and $\boldsymbol{c}_m$ agree that the pair belongs together, thus pair$(\boldsymbol{c}_z \times \boldsymbol{c}_m, \boldsymbol{c}_z) = g^z - g^{\text{agree}}$. By the same argument pair$(\boldsymbol{c}_z \times \boldsymbol{c}_m, \boldsymbol{c}_m) = g^m - g^{\text{agree}}$ and therefore pair$(\boldsymbol{c}_z \times \boldsymbol{c}_m, \boldsymbol{c}_m) +$ pair$(\boldsymbol{c}_z \times \boldsymbol{c}_m, \boldsymbol{c}_z) =$ pair$(\boldsymbol{c}_z, \boldsymbol{c}_m)$.

Thus, the meet provides a natural definition of the area between two clusterings, so we will require that $d(\boldsymbol{c}_z, \boldsymbol{c}_m) = d(\boldsymbol{c}_z \times \boldsymbol{c}_m, \boldsymbol{c}_z) + d(\boldsymbol{c}_z \times \boldsymbol{c}_m, \boldsymbol{c}_m)$. □

Lemma 3. *If clustering* $\boldsymbol{c}_j^{'}$ *refines* $\boldsymbol{c}_j$ *then pair*$(\boldsymbol{c}_j, \boldsymbol{c}_j^{'}) = \sum_{k=1}^{K_j}$ *pair*$(\boldsymbol{c}(1, n_{jk}), \boldsymbol{c}^{'}(n_{jk}))$

Proof. Define $K_{j'}$ as the number of clusters in the refined clustering. Apply the definition of the pair function results in pair$(\boldsymbol{c}_j, \boldsymbol{c}_j^{'}) = \sum_{k=1}^{K_j} \binom{n_{jk}}{2} - \sum_{z=1}^{K_{j'}} \binom{n_{j'z}}{2}$ (because the refinement can only break apart pairs). For each cluster k in $\boldsymbol{c}_j$, enumerate the clusters in $\boldsymbol{c}_j^{'}$ that refine k with $r = 1, \ldots, R_k$ (and note, R_k could be 1, indicating that there was no refinement). We can rewrite the pair function as pair$(\boldsymbol{c}_j, \boldsymbol{c}_j^{'}) = \sum_{k=1}^{K_j} \left(\binom{n_{jk}}{2} - \sum_{r=1}^{R_k} \binom{n_{j'r}}{2} \right) = \sum_{k=1}^{K_j}$ pair$(\boldsymbol{c}(1, n_{jk}), \boldsymbol{c}^{'}(n_{jk}))$. □

Theorem 1 (Meila, 2007)**.** *The three assumptions imply that the distance metric is the Variation of Information, given by*

$$d(\boldsymbol{c}_j, \boldsymbol{c}_{j'}) \equiv VI(\boldsymbol{c}_j, \boldsymbol{c}_{j'}) \quad = \quad H(\boldsymbol{c}_j|\boldsymbol{c}_{j'}) + H(\boldsymbol{c}_{j'}|\boldsymbol{c}_j). \tag{B.1}$$

Proof. The four properties, derived from our three assumptions are equivalent to those stated in Meila (2007), and so the proof follows the same argument, which we present here for completeness. Applying the third and fourth properties we see that $d(\boldsymbol{c}_j, \boldsymbol{c}(N,N)) = \sum_k^{K_j} \frac{n_k}{N} d(\boldsymbol{c}(1, N_k), \boldsymbol{c}(\text{uniform}(N_k), N_k)) = \sum_k^{K_j} \frac{n_k}{N} \log n_k$ Adding and subtracting $\log N$ we have $\sum_k^{K_j} \frac{n_k}{N} \log n_k = \sum_k^{K_j} \frac{n_k}{N}(\log \frac{n_k}{N} + \log N)$, which is equal to $\log N - \text{H}(\boldsymbol{c}_j)$. By our fourth property $d(\boldsymbol{c}(1,N), \boldsymbol{c}(N,N)) = \log N$. Property 1 and this fact imply $d(\boldsymbol{c}_j, \boldsymbol{c}(1,N)) = \text{H}(\boldsymbol{c}_j)$. Now, consider two arbitrary clusterings, $\boldsymbol{c}_m$ and $\boldsymbol{c}_z$. Identify all the n_{km} observations assigned to the k^{th} cluster in method m as k_m. And collect the cluster labels for these documents in $\boldsymbol{c}_z$ in $\boldsymbol{c}_z(k_m)$. Then, $d(\boldsymbol{c}_m, \boldsymbol{c}_m \times \boldsymbol{c}_z) = \sum_{k=1}^{K_m} \frac{n_{km}}{N} d(\boldsymbol{c}(1, n_{km}), \boldsymbol{c}_z(k_m))$ and by our previous argument $\sum_{k=1}^{K_m} \frac{n_{km}}{N} d(\boldsymbol{c}(1, n_{km}), \boldsymbol{c}_z(k_m)) = \sum_{k=1}^{K_m} \frac{n_{km}}{N} H(\boldsymbol{c}_z(k_m))$ and applying properties of entropy reveals that $\sum_{k=1}^{K_m} \frac{n_{km}}{N} H(\boldsymbol{c}_z(k_m)) = H(\boldsymbol{c}_m|\boldsymbol{c}_z)$. Applying our second property then shows that $d(\boldsymbol{c}_m, \boldsymbol{c}_z) = H(\boldsymbol{c}_m|\boldsymbol{c}_z) + H(\boldsymbol{c}_z|\boldsymbol{c}_m)$ which completes the proof. □

C The Sammon Multidimensional Scaling Algorithm

We define here the Sammon (1969) multidimensional scaling algorithm and shows that it possesses the properties we need. Let $\boldsymbol{c}_j$ be an $N \times K_j$ matrix (for document i, $i = 1, \ldots, N$, and cluster k, $k = 1, \ldots, K_j$, characterizing clustering j), each element of which describes whether each document is (0) or is not (1) assigned to each cluster (or for soft clustering methods how a document is allocated among the clusters, but where the sum over k is still 1). For each clustering j, the goal is to define its coordinates in a new two-dimensional space $\boldsymbol{x}_j = (x_{j1}, x_{j2})$, which we collect into a $J \times 2$ matrix $\boldsymbol{X}$. We use the Euclidean distance between two clusterings in this space, which we represent for clusterings j and $j^{'}$ as $d^{\text{euc}}(\boldsymbol{x}_j, \boldsymbol{x}_{j'})$. Our goal is to estimate the coordinates $\boldsymbol{X}^*$ that minimizes

$$\boldsymbol{X}^* \quad = \quad \texttt{argmin}_{\boldsymbol{X}} \left(\sum_{j=1}^{J} \sum_{j' \neq j} \frac{\left(d^{\text{euc}}(\boldsymbol{x}_j, \boldsymbol{x}_{j'}) - d(\boldsymbol{c}_j, \boldsymbol{c}_{j'}) \right)^2}{d(\boldsymbol{c}_j, \boldsymbol{c}_{j'})} \right). \tag{C.1}$$

Equation C.1 encodes our goal of preserving small distances with greater accuracy than larger distances. The denominator contains the distance between two clusterings $d(\boldsymbol{c}_j, \boldsymbol{c}_{j'})$. This implies that clusterings that are small will be given additional weight in the final embedding, while large distances will receive less consideration in the scaling, just as desired.

D The Local Cluster Ensemble

This appendix describes two properties of the local cluster ensemble, defined in Section 3.5.

Avoiding Infinite Regress via Local Cluster Ensembles We first show that the local cluster ensemble avoids the infinite regress problem. To prove this, we show that our approach is invariant to replacing the k-means "meta-cluster analysis" method used in Section 3.5 to form local cluster ensembles with any other valid clustering method, given that we employ a sufficiently large number of methods in the original set. Suppose that we employ any valid distance metric between clusterings and apply arbitrary clustering method 1 to obtain a partition of documents based upon the weighted votes for a given point. We represent this clustering with $\boldsymbol{c}_1(\boldsymbol{V}(\boldsymbol{w}))$. Now, suppose that we want to apply a second cluster method to same weighted voting matrix $\boldsymbol{c}_2(\boldsymbol{V}(\boldsymbol{w}))$. How close can we get to $\boldsymbol{c}_1(\boldsymbol{V}(\boldsymbol{w}))$ by varying the weights in $\boldsymbol{c}_2(\boldsymbol{V}(\boldsymbol{w}))$?[4] If it is close, then we are guaranteed to find the same clusterings (and therefore, make the same discoveries) using two different clustering methods.

Let $\boldsymbol{w}^*$ represent the set of weights that minimize the distance between $\boldsymbol{c}_1(\boldsymbol{V}(\boldsymbol{w}))$ and $\boldsymbol{c}_2(\boldsymbol{V}(\boldsymbol{w^*}))$, $\boldsymbol{w}^* = \texttt{argmin}_{\boldsymbol{w}'} d(\boldsymbol{c}_1(\boldsymbol{V}(\boldsymbol{w})), \boldsymbol{c}_2\left(\boldsymbol{V}(\boldsymbol{w}')\right)$. We can guarantee that, $0 \leq d(\boldsymbol{c}_1(\boldsymbol{V}(\boldsymbol{w})), \boldsymbol{c}_2(\boldsymbol{V}(\boldsymbol{w^*})) \leq \min_j d(\boldsymbol{c}_1(\boldsymbol{V}(\boldsymbol{w}), \boldsymbol{c}_j))$ or that $\boldsymbol{c}_1(\boldsymbol{V}(\boldsymbol{w}))$ and $\boldsymbol{c}_2(\boldsymbol{V}(\boldsymbol{w^*}))$ can be no farther apart $\boldsymbol{c}_1(\boldsymbol{V}(\boldsymbol{w}))$ and any of the clusterings we have already obtained. This is because we can always place all the weight on the clustering from an existing method. If we have included all possible clusterings for a set of documents, then $d(\boldsymbol{c}_1(\boldsymbol{V}(\boldsymbol{w})), \boldsymbol{c}_2(\boldsymbol{V}(\boldsymbol{w^*})) = 0$, because the clustering from $\boldsymbol{c}_1(\boldsymbol{V}(\boldsymbol{w}))$ is guaranteed to be present in the collection of clusterings.

This illustrates two key points about the invariance of our method to the clustering method used in creating local cluster ensembles. First, because we use a large number of clusterings to obtain

[4] We make the assumption that the second clustering method is full range (can provide any partition) to avoid pathological counter examples. For simplicity, we also assume that when provided with a similarity matrix that is block-diagonal (diagonal blocks are zero distance, off diagonal infinite distance) the method returns the block-diagonals as the clustering. We are unaware of any existing clustering methods that violate this assumption, although theoretical examples are possible to construct. Notice, that our assumptions are different than Kleinberg (2003), avoiding well-known impossibility results.

many partitions of the data, any two methods used to cluster the results are likely to yield very similar insights. Second, we recognize that we cannot enumerate all possible partitions. Therefore, we restrict our attention only to those partitions that can be expressed by a combination of the collective creativity of the various academic literatures devoted to cluster analysis.

The Local Cluster Ensemble as a Relaxed Version of the "Meet" We now show that the local cluster ensemble is a *relaxed* version of the "meet," defined in Appendix A: it agrees in specific cases where we would expect correspondence, and diverges to allow *local* averages. In comparison, the original version of the meet creates a cluster ensemble that gives each component method equal weight.

We first demonstrate that the meet and the local cluster ensemble agree in specific cases. Consider two clusterings $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$ and denote their meet by $\boldsymbol{c}_3 = \boldsymbol{c}_1 \times \boldsymbol{c}_2$. Recall that a pair of documents will be assigned to the same cluster in $\boldsymbol{c}_3$ if (and only if) they are assigned to the same cluster in $\boldsymbol{c}_1$ and $\boldsymbol{c}_2$. To construct the meet using a local cluster ensemble, suppose that we assign equal weight to each method $w_1 = w_2 = 0.5$ and that the local cluster ensemble assumes the same number of clusters as found in the meet, K_3. A consequence of these assumptions is that pairs of documents assigned to the same cluster in both documents will be maximally similar. The optimal solution for k-means, applied to this similarity matrix is the meet (anything else will increase the squared error in the final clustering, and therefore not be optimal). Further, it is clear that the meet of a set of clusterings provides an upper bound on the number of clusters to be found in an ensemble: using more clusters than the meet involves splitting pairs of documents that are maximally similar into *different* clusters.

We now show how the meet relates to the local cluster ensemble in general. The meet among a set of clusterings requires unanimous agreement that a pair of documents belongs to the same cluster (the order of the pairs is irrelevant). We show this explicitly in terms of a voting matrix to compare it more directly to the local cluster ensemble. Suppose we have J clusterings and assemble the voting matrix $\boldsymbol{V}(\boldsymbol{w})$, but suppose each clustering receives equal weight $w = \frac{1}{J}$ and obtain similarity matrix $\boldsymbol{V}(\boldsymbol{w})\boldsymbol{V}(\boldsymbol{w})'$. The meet settles disputes about which documents belong in the same clusters in the most conservative way possible: requiring unanimous agreement among the clusterings that the pairs belong together.

Rather than require unanimous agreement among all clusterings to place a pair of documents in the same cluster — which would result in highly fragmented clusterings — the local cluster ensemble employs a non-unanimous voting rule; this allows for some clusterings to exert greater influence through arbitrary weights across the methods encoded in the vote matrix $\boldsymbol{V}(\boldsymbol{w})$. We then tally the total votes for each pair belonging together with $\boldsymbol{V}(\boldsymbol{w})\boldsymbol{V}(\boldsymbol{w})'$. The meta-clustering algorithm then adjudicates disputes among the clusterings about which documents belong together.

E Efficiently Sampling for Cluster Quality

Here we prove that if two clusterings agree about a pair of documents — both clusterings placing the pair together in a cluster or separately in different clusters — then it does not contribute to differences in our measure of cluster quality and so resources need not be devoted to evaluating it. Our evaluation then only needs to address pairs for which clusterings disagree. Define Y as 1 if the clusterings agree about a pair and 0 if they disagree, π_a as the proportion of pairs that agree, and $1 - \pi_a = \pi_d$ as the proportion of pairs that disagree. Then,

$$\begin{aligned} \mathrm{E}[\mathrm{CQ}(\boldsymbol{c}_j - \boldsymbol{c}_{j'})] &= \mathrm{E}\left[\mathrm{E}[\mathrm{CQ}(\boldsymbol{c}_j - \boldsymbol{c}_{j'})|Y]\right] \\ &= \pi_a \underbrace{\mathrm{E}[\mathrm{CQ}(\boldsymbol{c}_j - \boldsymbol{c}_{j'})|Y=1]}_{0} + \pi_d \underbrace{\mathrm{E}[\mathrm{CQ}(\boldsymbol{c}_j - \boldsymbol{c}_{j'})|Y=0]}_{\text{Estimated by Sampling}} \end{aligned} \tag{E.1}$$

The only piece of Equation E.1 that is unknown is the average cluster quality among the pairs where the two clusterings disagree. We can obtain an unbiased estimate of this by randomly sampling from the pairs where the two methods disagree and then obtain an unbiased estimate of the difference in cluster quality by multiplying by the proportion of pairs where there is disagreement π_d (which is easily computed from the population of pairs).

References

Abelson, H., K. Ledeen and H. Lewis. 2008. "Blown to bits: your life, liberty, and happiness after the digital explosion.".

Armstrong, J.S. 1967. "Derivation of theory by means of factor analysis or Tom Swift and his electric factor analysis machine." *American Statistician* pp. 17–21.

Bailey, Kenneth D. 1994. *Typologies and taxonomies: an introduction to classification techniques.* Beverly Hills: Sage.

Banerjee, Arindam, Inderjit Dhillon, Joydeep Ghosh and Suvrit Sra. 2005. "Clustering on the Unit Hypersphere Using von Mises-Fisher Distributions." *Journal of Machine Learning* 6:1345–1382.

Blei, David M., Andrew Y. Ng and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3:993–1022.

Blei, David and Michael Jordan. 2006. "Variational Inference for Dirichlet Process Mixtures." *Journal of Bayesian Analysis* 1(1):121–144.

Caruana, Rich, Mohamed Elhawary, Nam Nguyen and Casey Smith. 2006. Meta clustering. In *ICDM'06. Sixth International Conference on Data Mining.* pp. 107–118.

Cowan, Nelson. 2000. "The Magical Number 4 in Short Term Memory: A Reconsideration of Mental Storage Capacity." *Behavioral and Brain Sciences* 24:87–185.

Dhillon, Inderjit. 2003. "Co-clustering Documents and Words Using Bipartite Spectral Graph Partitioning." *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* pp. 89–98.

Diaconis, Persi, Sharad Goel and Susan Holmes. 2008. "Horseshoes in multidimensional scaling and local kernel methods." *Annals of Applied Statistics* 2(3):777–807.

Elman, Colin. 2005. "Explanatory Typologies in Qualitative Studies of International Politics." *International Organization* 59(2, spring):293–326.

Eulau, Heiz and Paul Karps. 1977. "The Puzzle of Representation: Specifying Components of Responsiveness." *Legislative Studies Quarterly* 2(3):233–254.

Fern, Xiaoli and Carla Brodley. 2003. Random Project for High Dimensional Data Clustering: A Cluster Ensemble Approach. In *Proceedings of the Twentieth International Conference on Machine Learning.*

Fiorina, Morris. 1989. *Congress, Keystone of the Washington Establishment.* New Haven: Yale University Press.

Fraley, C. and A.E. Raftery. 2002. "Model-based clustering, discriminant analysis, and density estimation." *Journal of the American Statistical Association* 97(458):611–631.

Frey, BJ and D Dueck. 2007. "Clustering by Passing Messages Between Data Points." *Science*

315(5814):972.

Gan, Guojun, Chaoqun Ma and Jianhong Wu. 2007. *Data Clustering: Theory, Algorithms, and Applications*. Philadelphia: Siam.

Gionis, A, H Mannila and P Tsaparas. 2005. Clustering aggregation. In *Proceedings of the 21st International Conference on Data Mining*.

Grimmer, Justin. 2009. "A Bayesian Hierarchical Topic Model for Political Texts: Measuring Expressed Agendas in Senate Press Releases." Harvard University.

Gutmann, Amy and Dennis Thompson. 1996. *Democracy and Disagreement*. Harvard University Press: Harvard University Press.

Guttman, L. 1950. "The problem of attitude and opinion measurement." *Measurement and prediction* 4.

Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2001. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.

Ho, Y and D Pepyne. 2002. "Simple Explanation of the No Free Lunch Theorem and Its Implications." *Journal of Optimization Theory and Applications* 115(3):549–570.

Hopkins, Daniel and Gary King. 2009, forthcoming. "A Method of Automated Nonparametric Content Analysis for Social Science." *American Journal of Political Science* . http://gking.harvard.edu/files/abs/words-abs.shtml.

Jones, Bryan, John Wilkerson and Frank Baumgartner. 2009. "The Policy Agendas Project.". http://www.policyagendas.org.

Kaufman, Leonard and Peter Rousseeuw. 1990. *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: Wiley.

King, Gary, Christopher J.L. Murray, Joshua A. Salomon and Ajay Tandon. 2004. "Enhancing the Validity and Cross-cultural Comparability of Measurement in Survey Research." *American Political Science Review* 98(1, February):191–207. http://gking.harvard.edu/files/abs/vign-abs.shtml.

King, Gary, Kay Schlozman and Norman Nie, eds. 2009. *The Future of Political Science: 100 Perspectives*. New York: Routledge Press.

Kleinberg, Jon. 2003. An Impossibility Theorem for Clustering. In *Advances in Neural Information Processing Systems Proceedings of the 2002 Conference*. pp. 463–470.

Kohonen, Teuvo. 2001. *Self-Organizing Maps.* New York: Springer.

Law, Martin, Alexander Topchy and Anil Jain. 2004. Multi-objective Data Clustering. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition.*

Lazardsfeld, Paul and Allen Barton. 1965. Qualitative Measurement in the Social Sciences: Classification, Typologies, and Indices. In *The Policy Sciences*, ed. Daniel Lerner and Harold Lasswell. Standard University Press.

Lewis, David. 1999. "Reuters -21578 text Categorization Test Collection Distribution 1.0.".

Mackay, David. 2003. *Information Theory, Inference, and Learning Algorithms.* Cambridge: Cambridge University Press.

Manning, Christopher D., Prabhakar Raghavan and Hinrich Schütze. 2008. *Introduction to Information Retrieval.* NY: Cambridge University Press.

Mayhew, D. 1974. "The electoral connection." *New Haven: Yale University* .

Meila, Marina. 2007. "Comparing Clusterings: An Information Based Distance." *Journal of Multivariate Analysis* 98(5):873–895.

Mikhaylov, Slava, Michael Laver and Kenneth Benoit. 2008. "Coder Reliability and Misclassification in Comparative Manifesto Project Codings." Paper presented at the Midwest Political Science Association, Chicago.

Miller, George. 1956. "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information." *Psychological Review* 63:81–97.

Miller, W.E. and D.E. Stokes. 1963. "Constituency influence in Congress." *The American Political Science Review* pp. 45–56.

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

Ng, Andrew, Michael Jordan and Yair Weiss. 2002. "On Spectral Clustering: Analysis and an Algorithm." *Advances in Neural Information Processing Systems 14: Proceedings of the 2002 Conference* .

Pitkin, Hanna F. 1972. *The Concept of Representation.* University of California Press.

Pitman, Jim. 1997. "Some Probabilistic Aspects of Set Partitions." *The American Mathematical Monthly* pp. 201–209.

Quinn, K.M., B.L. Monroe, M. Colaresi, M.H. Crespin and D.R. Radev. 2006. "How To Analyze Political Attention With Minimal Assumptions And Costs." Annual Meeting of the Society for Political Methodology.

Sammon, John. 1969. "A Nonlinear Mapping for Data Structure Analysis." *IEEE Transactions on Computers* 18(5):401–409.

Schrodt, P.A. and D.J. Gerner. 1997. "Empirical indicators of crisis phase in the Middle East, 1979-1995." *Journal of Conflict Resolution* pp. 529–552.

Shawe-Taylor, John and Nello Cristianini. 2004. *Kernel Methods for Pattern Analysis*. Cambridge: Cambridge University Press.

Shi, J and J Malik. 2000. "Normalized Cuts and Image Segmentation." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(8):888–905.

Simon, Herbert. 1957. *Models of Man*. New York: Wiley.

Spivey, M.Z. 2008. "A generalized recurrence for Bell numbers." *J. Integer Sequences* 11.

Strehl, Alexander and Joydeep Grosh. 2002. "Cluster Ensembles: A Knowledge Reuse Framework for Combining Multiple Partitions." *Journal of Machine Learning Research* 3:583–617.

Teh, Y, M Jordan, M Beal and D Blei. 2006. "Hierarchical Dirichlet Processes." *Journal of the American Statistical Association* 101(476):1566–1581.

Topchy, A, AK Jain and W Punch. 2003. Combining Multiple Weak Clusterings. In *Proceedings IEEE International Conference on Data Mining*.

von Luxburg, Ulrike. 2007. "A Tutorial on Spectral Clustering." *Statistics and Computing* 17(4):395–416.

Watanabe, Satosi. 1969. *Knowing and Guessing: A Quantitative Study of Inference and Information*. New York: Wiley.

Wolpert, DH and WG Macready. 1997. "No Free Lunch Theorems for Optimization." *IEEE Transactions on Evolutionary Computation* 1(1):67–82.

Yiannakis, Diane Evans. 1982. "House Members Communication Styles: Newsletters and Press Releases." *Journal of Politics* 44(4):1049–1071.