The Migration of the Field Code

Kellen Funk and Lincoln Mullen

February 17, 2016

This handout describes our work in progress for the Digital Humanities Working Group at George Mason University.

E-mail: `lincoln@lincolnmullen.com`, `kfunk@princeton.edu`

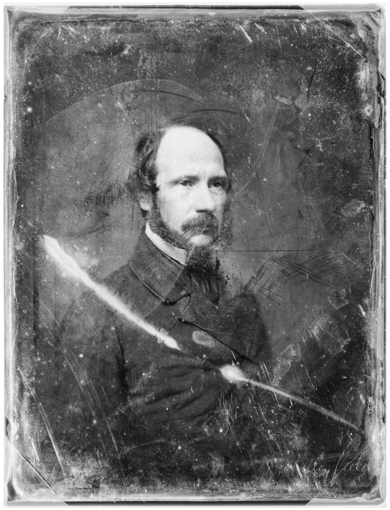

Figure 1: David Dudley Field II (1805–1894). Daguerreotype by Matthew Brady, circa 1844 to 1860, DAG no. 084, Library of Congress, Prints and Photographs Division, Washington, DC: `http://www.loc.gov/pictures/item/2004663945/`.

The Field Code

After the American Revolution, most states were common law jurisdictions, sometimes with courts of chancery. These courts had a complex system of pleading defined by case law. By the 1840s, lawyers and the mercantile classes called for the simplification and rationalization of civil procedure through codification. Factions for and against codification debated about whether codes were laws passed by democratic legislatures or anti-democratic legal elites, whether codification served only the needs of wealthy capitalists, and whether the purity of Anglo-Saxon civilization derived from the common law would be maintained. The economic capital New York was the first state to codify its procedure in 1848, thanks to the efforts of David Dudley Field. By the end of the century, New York's Field Code became the model for the codes of civil procedure in most states.

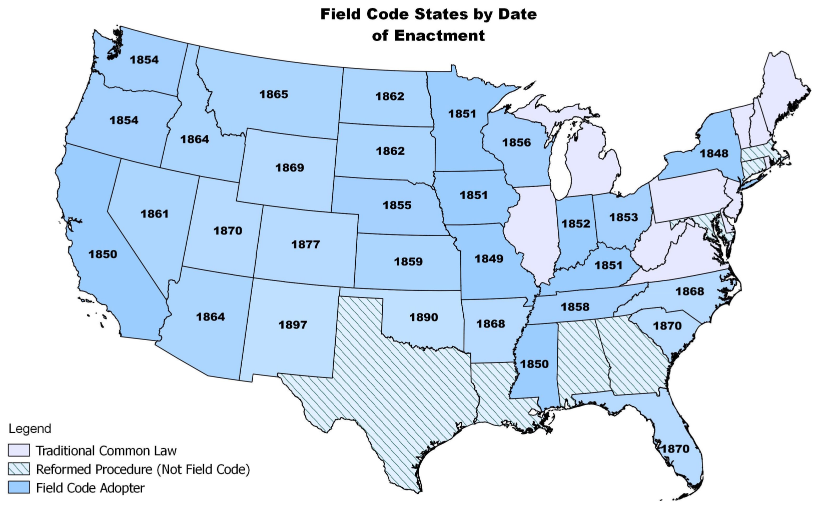

Figure 2: Field Code states by date of first enactment. Many states subsequently revised their codes of civil procedure.

Legal historians have long known that the Field Code spread to other jurisdictions. Beyond the mere fact of its adoption, however, no one has studied the content of the borrowings. Which codes borrowed from each other? Which sections were borrowed, and how were they modified? What were the patterns and structures of borrowings and of innovations? To answer these questions we gathered a corpus of 115 codes and statutes of civil procedure from 1806 to 1933 containing about 7.6 million words, then algorithmically detected the borrowings within that corpus.

The aim of this handout is to describe our contributions to digital methods in legal history (and other fields), and to outline the interpretations we have drawn about the migration of the Field Code. Our method is one of two common approaches in computational history. We created a dataset to answer a given set of questions; a different approach is to take the sources as given and explore the data to see what questions it raises.[1] Our historical findings describe how American legal practice became standardized around a New York code yet varied by region.

How we found the borrowings

We found out how the codes borrowed from one another by splitting the codes into sections and comparing each section to every other section.[2] This process in essence mimicked the way that nineteenth-century code commissioners literally cut and pasted sections from other jurisidictions.

Preparing the corpus

Having identified all of the relevant laws of civil procedure in the nineteenth century, including codes, session laws, and statutory compilations, we used OCR software to create plain-text versions of the codes. These OCR files received only a light cleaning: we edited the section markers by hand as necessary, and wrote a script to fix the most obvious OCR errors.

We then split each section of each code into its own text file.[3] The corpus contains nearly 98,000 sections. Below is a sample file containing a single section from a single code.[4]

```
151. An issue arises when a fact or conclusion of law
is maintained by the one party, and is controverted
by theothsr. Issues are of two kinds: lst. Of law:
and,  _ 2d. or fact. if?
```

Tokenizing and measuring similarity

Next we tokenized the text into shingled n-grams.[5] After experimenting we found that five-grams worked best for detecting similarity

[1] This approach features "middle data" (not "big data") which we might define as data that is too small for distributed computation but too big for naive algorithms. Alternatively, it is data where the size of the sample approaches the size of the population, but where the population is strictly constrained by the research problem.

[2] The data and code to re-run our analyses are available in a GitHub repository: `https://github.com/lmullen/civil-procedure-codes`. We generalized our method in Lincoln Mullen, "textreuse: Detect Text Reuse and Document Similarity," R package version 0.1.2 (2015): `https://github.com/ropensci/textreuse`, which was peer-reviewed by rOpenSci. We benefitted a great deal from David A. Smith, Ryan Cordell, and Elizabeth Maddock Dillon, "Infectious Texts: Modeling Text Reuse in Nineteenth-Century Newspapers," in *2013 IEEE International Conference on Big Data*, 2013, 86–94, `doi:10.1109/BigData.2013.6691675`; David A. Smith et al., "Detecting and Modeling Local Text Reuse," in *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries* (IEEE Press, 2014), 183–92, `http://dl.acm.org/citation.cfm?id=2740800`.

[3] While all statutes are divided into chapters, articles, and titles, specific regulations appear within these as sequentially numbered sections.

[4] 1851 California Laws 74, §151. This corresponds to our file `CA1851-001660.txt`. Notice the minor OCR errors.

[5] We also hashed the tokens, meaning that we converted them to integer representations.

despite intentional word changes and OCR errors. Below are the first five tokens from the section above.

```
[1] "151 an issue arises when"
[2] "an issue arises when a"
[3] "issue arises when a fact"
[4] "arises when a fact or"
[5] "when a fact or conclusion"
```

Next we used the Jaccard similarity score for measuring document similarity. That measure, treating the tokens as a set, is the ratio of shared tokens to the ratio of total tokens in two documents. That ratio can range from 0 (complete dissimilarity) to 1 (complete similarity).[6] For instance, the section from the 1851 California code above was derived from the 1850 New York code.[7] Because of changes to the wording and OCR errors, the Jaccard similarity between the two sections was 0.189. By carefully checking matching sections versus scores, we arrived at a rule of thumb that a Jaccard similarity score greater than 0.15 likely indicated a match, and a score greater than 0.2 almost certainly indicated a match.[8]

Computing similarity for the entire corpus

The next step was to compute the similarity of every section in the corpus to every other section. That posed a problem: doing so would require an enormous number of comparisons, approximately 4.8 billion for our corpus.[9] Most of these comparisons would be unnecessary, since each section has no relationship to most other sections.

We implemented the minhash/locality sensitive hashing (LSH) algorithm to detect candidate pairs, i.e., documents which were likely to be matches. This algorithm works by extracting a set number of random tokens from each document, allowing the documents to be represented uniformly and compactly. Then those random tokens are grouped into subsets. If any two documents have a matching subset, then they are considered a candidate pair. This algorithm has several useful properties. It approximates the Jaccard similarity of the two documents. It requires a computation for each document rather than each pair of documents, so the computation time grows linearly not geometrically. And by controlling various parameters, one can set a threshold similarity score above which one is likely to find a genuine match and unlikely to find a spurious match.

Once we detected the candidate pairs, we measured the actual Jaccard similarity of those pairs. The result was a sparse matrix of similarity scores, with rows and columns for each section in the corpus.

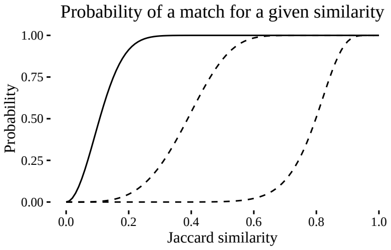

Figure 3: This chart shows the threshold S-curves for various settings of the minhash/LSH algorithm. The x-axis shows the actual measured Jaccard similarity of the two documents; the y-axis shows the probability that they will be marked as a match. We used the settings for the leftmost curve, guaranteeing that we detected all matches above a similarity of 0.2.

[6] Jure Leskovec, Anand Rajaraman, and Jeff Ullman, *Mining of Massive Datasets*, 2nd ed. (Cambridge University Press, 2014), ch. 3, `http://www.mmds.org`. The formal definition of the Jaccard similarity score is

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

[7] *Final Report of the Commission on Practice and Pleadings*, 2 *Documents of the Assembly of New York*, 73rd session, No. 16 (1850), 317, §756. This corresponds to our file `NY1850-008350.txt`.

[8] The mean score for the best section matches was 0.54.

[9] Assuming that the similarity measure is bi-directional, the number of pairwise comparisons in a corpus is given by $(n^2 - n)/2$.

Table 1: A subset of the filtered similarity matrix.

	NY1850-018680	CA1851-004380	MN1851-010470	OR1854-003380	WA1855-003150
NY1850-018680					
CA1851-004380	0.27				
MN1851-010470	0.26				
OR1854-003380	0.39				
WA1855-003150				0.47	

This matrix required filtering based on what we knew about the process of borrowing. For instance, a code from 1851 obviously did not borrow from a code from 1877. Furthermore, in chains of borrowing (e.g., NY1850 → CA1851 → CA1868 → CA1872 → MT1895) the latest section might have a high similarity to all of its parents, but it was in fact borrowed only from the most recent parent. We therefore filtered the similarity matrix to remove matches within the same code; anachronistic matches; spurious matches beneath a certain threshold. Then if a section had multiple matches, we kept the match from the chronologically closest code, giving preference to codes from the same state, unless there was a substantially higher match from a different code. The result was a sparse matrix of the most likely matches for each section, with at most one match for each section.

Read this table as `row` borrows from `column`, e.g., `CA1851-004380` borrows from `NY1850-018680`.

Learning from the borrowings

A similarity matrix is a common input to many algorithms and visualizations. From that matrix, we built several means of understanding how the codes borrowed from one another.

Clustering the borrowings

We used a clustering algorithm to group related sections together. There are innumerable clustering algorithms, but we needed one that could work with a sparse matrix and one whose assumptions matched the characteristics of our problem. We decided to use affinity propagation clustering.[10] That algorithm assumes that each cluster has an "exemplar" item which the other items are similar to. That assumption fits nicely with borrowings from the Field Code, where a single section (likely from the 1850 New York code) had many borrowings. Furthermore, even though the affinity propagation clustering algorithm did not fully converge with our peculiar dataset, it did an adequate job clustering the documents. Because there was

[10] Brendan J. Frey and Delbert Dueck, "Clustering by Passing Messages Between Data Points," *Science* 315 (2007): 972–976.

an exemplar section for each cluster, we were able to merge clusters whose exemplars had a high Jaccard similarity score.

The result was a set of approximately 2,900 clusters which contained at least five sections, though this probably overstates the number of ur-sections in the corpus. Each cluster contained a list of the section that belonged to it. The biggest cluster, for instance, which concerned the use of affidavits in pleading, contained 103 sections. Scholars in digital literary studies often speak of the "deformance" of texts.[11] Clustering the sections of the codes deformed them by pulling the sections out of their context within specific codes and by arranging them by topic in chronological order. Deforming the codes allowed us to see the development and spread of the law. Consider this excerpt from a cluster of sections about the competence of witnesses:

```
Cluster ID: 10382          Documents in cluster: 20
Exemplar: OR1854-003380    Earliest: NY1850-018680

NY1850-018680 -----------------------------------------------------------------
1709. The following persons are not admissible: 1. Those who are of unsound mind
at the time of their production for examination: 2. Children under ten years of
age, who appear incapable of receiving just impressions of the facts, respecting
which they are examined, or of relating them truly.

CA1851-004380 -----------------------------------------------------------------
394. The following persons shall not be witnesses: ‘ lst. Those who areof
unsound mind at the time of their production for examination: 2d. Children under
ten years of age, who appear incapable of receiving just impressions of the
facts respecting which they are examined, or of relating them truly: and, 3d.
Indians, or persons having one fourth or more of Indian blood, in an action or
proceeding to which a white person is a party: 4th. Negroes, or persons having
one half or more Negro blood, in an action or proceeding to which a white person
is a party.

OR1854-003380 -----------------------------------------------------------------
6. The following persons shall not be competent to testify 1. Those who are of
unsound mind, or intoxicated at the time of their production for examination ;
2. Children under ten years of age, who appear incapable of receiving just
impressions of the facts respecting which they are examined, or of relating them
truly; 4. Negroes, mulattocs and Indians, or persons one half or more of Indian
blood, in an action or proceeding to which a white person is a party.
```

Reading through those chronologically arranged sections, it became apparent how the original New York section was borrowed but adapted to include several different versions of a racial exclusion

[11] Stephen Ramsay, *Reading Machines: Toward an Algorithmic Criticism* (University of Illinois Press, 2011), ch. 3.

based on blood quantum.

Networks of borrowings

A matrix of similarities can also be thought of as the adjacency matrix of a network graph. Instead of creating a network graph of section to section borrowings (which would be practically the same thing as clustering), we moved up a level of abstraction to create a network graph of code to code borrowings. Because even our efforts at determining the best match for each section sometimes attributed a section to an incorrect code, we pruned the edges of the graph so that each code was connected to another code only if it borrowed at least fifty sections or twenty percent of its sections.

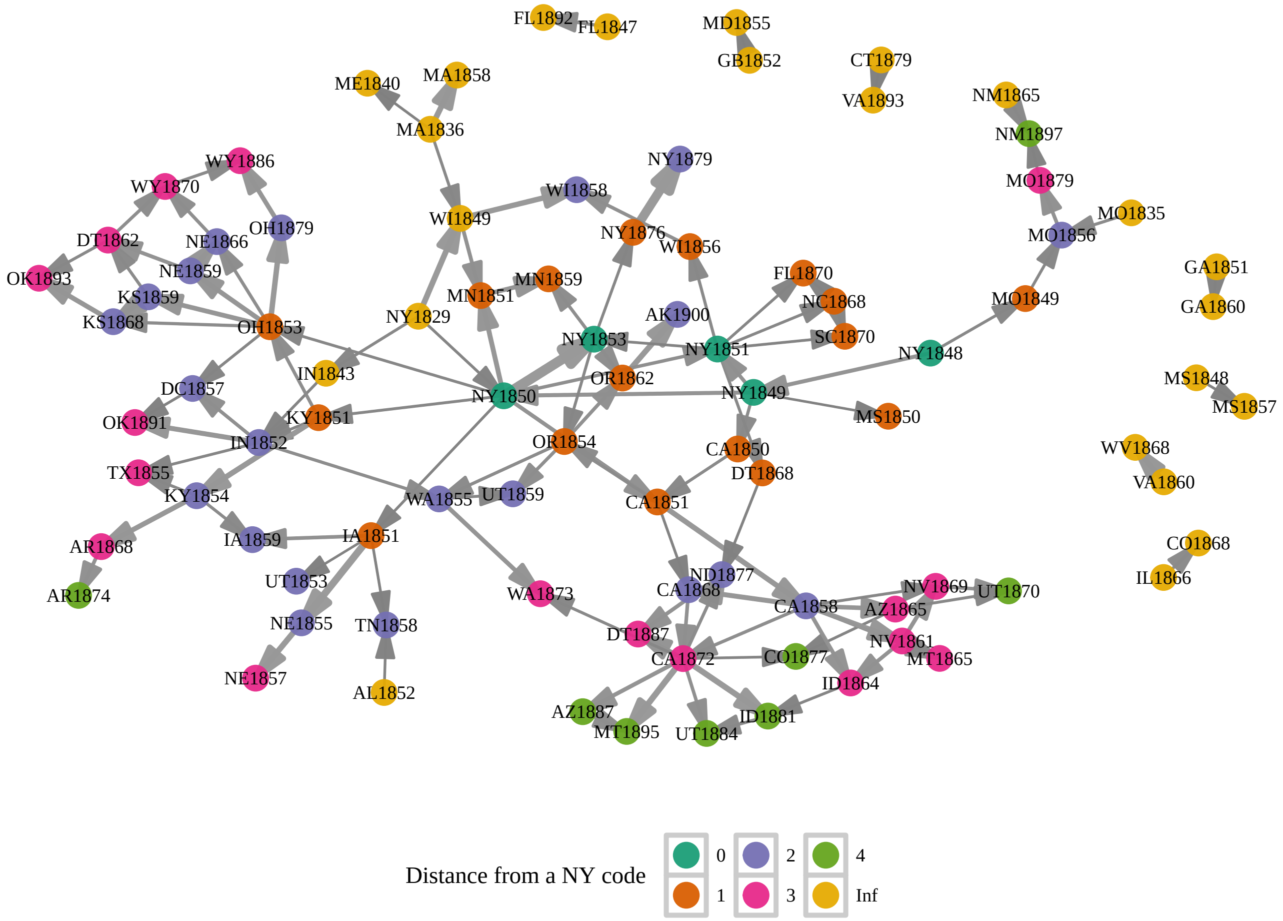

Figure 4: Code to code borrowings. The arrows indicate that a code contributed sections to another code. The color of the nodes indicates the number of steps to get to one of the New York codes. We could move up a further level of abstraction to consider state to state borrowings, which would show us the structure more clearly, albeit at a considerable loss of detail.

This network shows the structure of borrowings within the corpus. This network graph expands both geographically an chronologically. Later codes tend to be on the outside of the diagram, and they are the most distant from the original Field Code in both time and similarity. The New York codes, especially the eponymous Field Code of 1850, are central to the entire network. The Field Code became the basis of other families of codes, which adapted it to different circumstances. Those families of codes tended to be regional, and the ties within the families of codes are somewhat tighter than the ties of the regional exemplar to the Field Code. Yet state code commissioners could and did sometimes borrow from multiple states.

Borrowings within each code

Finally, we created a "spectrograph" visualization in order to get a fuller picture of how each code compiled sections from other codes. These visualizations showed the origin of each section of a code. We can use these spectrographs to follow the some of the connected nodes on the network diagram. Take Washington's 1855 code.

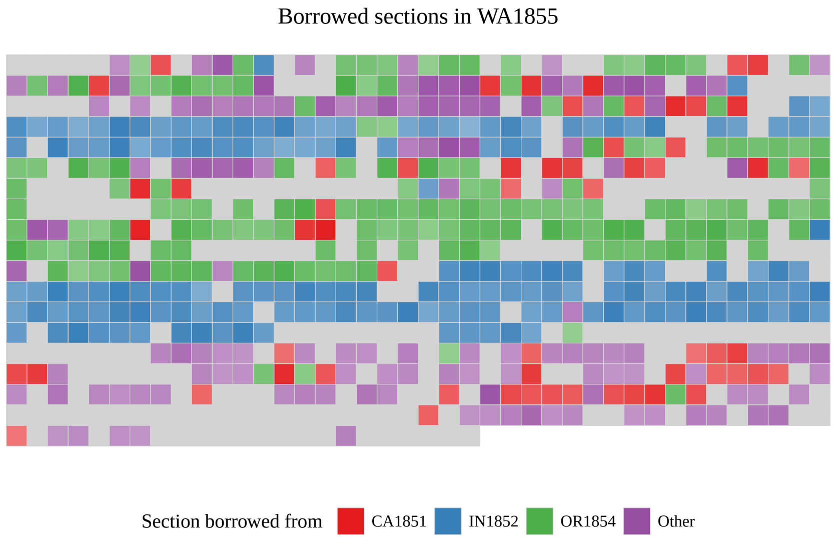

Figure 5: Borrowed sections in Washington's 1855 code. Each section in the code is represented by a square, the color of which indicates where it was borrowed from.

While there are a number of "other" sections, the pattern of borrowing is clear: Indiana's 1852 code and Oregon's 1854 code provide the majority of the borrowings. Those two bands of borrowings correspond to regulations on judgement from Oregon and enforcement provisions from Indiana. We think this pattern is because one of the Washington code commissioners, Edward Lander, was an Indiana appelate judge from 1850 to 1853, while another commissioner, William Strong, was a justice of the Oregon Supreme Court in the same years. The law was created by the commissioners, and they picked the laws that they knew best.

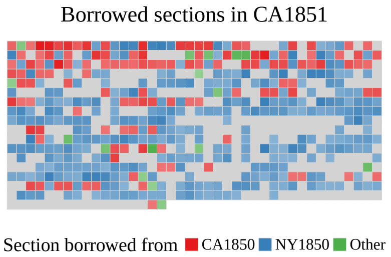

Figure 6: Borrowed sections in California's 1851 code. This code, which is the regional exemplar for much of the western codes, is almost entirely derived from the Field Code, with the exception of sections modified for California in its 1850 code. Those too were ultimately derived from New York.

Conclusions about method and interpretation

By framing our historical questions around "medium data" we enjoyed a useful symbiosis of traditional and digital historical methods. Our computational methods produced useful historical knowledge because they were justified by what we knew about the data from traditional historical work. We knew that code commissioners worked with "the scissors and paste-pot,"[12] as political debates from the era frequently complained, and we knew of codes in the archives that commisions literally marked up the legislation of other states.[13] Knowing that borrowings happened section by section justified our use of the minhash/LSH algorithm, and the nature of the borrowings justified our use of affinity propagration clustering. And knowing how codes passed from state to state made network analysis an obvious fit.[14] Our method is applicable to many historical questions and sources, especially corpora where the documents can be readily divided into sections. We may pursue it with other legal documents, such as treatises or statutes, or with sources for religious history, such as hymnbooks or tracts.

We are writing an article with a fuller statement of our historical interpretations, but to sum up what we learned from the computational methods: This study is significant because it gets at the heart of lawmaking in U.S. history. Lawyers and judges, politicians and newspaper editors warred over whether codes that were drafted by commissioners and borrowed wholesale from one another were actually democratic laws or the imposition of a legal elite. Furthermore, these borrowings called into question the extent to which the United States was actually a federal system of laws. We have shown the dominance of New York's laws in a supposedly federal system, and demonstrated the extent to which code commissioners borrowed from the Field Code and its descendants. Despite the notion that states were equal and sovereign, economic centers were in fact legal centers. But we have also demonstrated that regional distinctions were also significant, with states like California and Ohio creating separate Western and Midwestern families of codes. Finally we have shown how haphazard the process of codification could be, as in the case of the 1855 Washington code cited above, or in the case of New Mexico, which borrowed Missouri's laws because those were the law books available to the commissioners. In most jurisdictions the exact language of the Field Code is not currently on the books, but its basic provisions for civil procedure are in force throughout the United States.[15] Without too much exaggeration we might say that our method has revealed the spine of modern American legal practice.

[12] *Rocky Mountain News*, January 20, 1877.

[13] The Nevada code was created by marking up a printed copy of California's code. Detail from Council Bill 21, First Territorial Legislative Session (1861), Nevada State Library, Archives and Public Records.

[14] This is an example of the "no free lunch theorem" in action. See David Robinson, "K-means clustering is not a free lunch," *Variance Explained*, January 16, 2015: `http://varianceexplained.org/r/kmeans-free-lunch/`, citing David H. Wolpert and William G. Macready, "No Free Lunch Theorems for Optimization" *IEEE Transactions on Evolutionary Computation* 1, no. 1 (April 1997): 67–82, `http://ti.arc.nasa.gov/m/profile/dhw/papers/78.pdf`.

[15] But in some states such as Missouri the Field Code provisions survive nearly unchanged.